IMPROVING SERVICE BY INFORMING CUSTOMERS ABOUT ANTICIPATED DELAYS

by

Ward Whitt[1]

AT&T Labs–Research

April 14, 1998

Management Science 45 (1999) 192–207

[1]Room A117, AT&T Labs, 180 Park Avenue, Florham Park, NJ 07932-0971; wow@research.att.com

Abstract

This paper studies alternative ways to manage a multi-server system such as a telephone call center. Three alternatives can be described succinctly by: (i) blocking, (ii) reneging and (iii) balking. The first alternative – blocking – is to have no provision for waiting. The second alternative is to allow waiting, but neither inform customers about anticipated delays nor provide state information to allow arriving customers to predict delays. The second alternative tends to yield higher server utilizations. The first alternative tends to reduce to the second, without the first-come first-served service discipline, when customers can easily retry, as with automatic redialers in telephone access. The third alternative is to both allow waiting and inform customers about anticipated delays. The third alternative tends to cause balking when all servers are busy (abandonment upon arrival) instead of reneging (abandonment after waiting). Birth-and-death process models are proposed to describe the performance with each alternative. Algorithms are developed to compute the conditional distributions of the time to receive service and the time to renege given each outcome. Algorithms are also developed to help the service provider predict customer waiting times before beginning service, given estimated service-time distributions and the elapsed service times of the customers in service. Better predictions may be obtained by classifying customers and thereby obtaining better estimates of their service-time distributions.

Keywords: service systems, telephone call centers, blocking, balking, reneging, retrials, abandonments, birth-and-death processes, predicting delays, communicating anticipated delays

1. Introduction

[sec1] In this paper we investigate alternative ways to manage a service system. We have in mind a telephone call center staffed by a group of operators, but there are other possible applications, e.g., internet access. We introduce birth-and-death (BD) stochastic process models that can be used to demonstrate the advantage of: (1) allowing waiting before beginning service and (2) communicating anticipated delays to customers upon arrival (or providing state information to allow customers to predict delays). However, if the service provider decides to inform customers about anticipated delays, then it is important for the service provider to be able to accurately predict the delays. Hence, we also propose methods for the service provider to estimate the delay distribution and its mean, exploiting estimated service-time distributions (which need not be exponential) and the elapsed service times of customers in service.

The frame of reference is the classical loss system, in which there are s servers working in parallel and no extra waiting space. Assuming that blocked customers depart without affecting future arrivals, allowing waiting helps to avoid blocking and thus serve more customers. However, in many loss systems with telephone access, blocked customers can easily retry rapidly because of automatic redialers. When customers can easily retry, the system without a provision for waiting tends to behave like the system with a provision for waiting. However, we contend that it is usually better to directly allow for waiting. An advantage of customers retrying is that more customers receive service, while the service provider avoids the cost of maintaining a queue. However, retrying typically imposes costs on both the customers and the service provider. First, the customer must expend time and effort retrying. Second, even unsuccessful attempts often consume resources of the service provider. Typically, some resources are required to process each request for service, whether or not it is successful. Thus, the service provider's processing capacity may be reduced by having to handle many unsuccessful attempts. Moreover, with retrials, the first-come first-served (FCFS) service discipline is lost. The FCFS discipline is often strongly preferred by customers because of its inherent fairness. The random order of service associated with retrials also makes the waiting time before beginning service more variable, which tends to be detrimental. Thus, there are several reasons motivating service providers to directly allow for waiting.

Given that the service provider allows for waiting, there are two alternatives. The service provider may either communicate anticipated delays to customers upon arrival or not. We contend that, once the service provider has decided to allow for waiting, it is usually much better to inform customers about anticipated delays, assuming that there is the capability of doing so, which is more and more becoming the case; e.g., see Rappaport [21]. The most convincing argument perhaps come from our feelings about our own experience as a customer.

If the service provider does communicate anticipated delays, then the customers are more likely to balk when all servers are busy (leave immediately upon arrival) than renege (leave after waiting for some time). We develop BD models to describe and compare these alternatives. It is common practice to restrict attention to the special case of the M/M/s/r model, which has s servers and r extra waiting spaces. Indeed, it is common to use only the Erlang B (loss) model ($r = 0$) or the Erlang C (delay) model ($r = \infty$), but none of these alternatives account for balking or reneging. However, it is actually not difficult to account for balking and reneging in a BD model, and it is often very important to do so. By having a BD model that incorporates all the possibilities, it is easy to evaluate the alternatives. The way to calculate the steady-state distribution of a general BD process is quite well known. We go beyond that initial step by showing how to compute the probability that a customer receives service, the probability that a customer reneges, and the distributions and first two moments of the conditional response time given that service is completed and the conditional time to renege given that the customer reneges. These descriptions are helpful because the conditioning can have a big impact.

These general BD models also can be used to study complex *networks* of service facilities. As in Whitt [28], Kelly [18] and Ross [23], the BD model can serve as the fundamental building block for a reduced-load approximation for a network of service facilities. Then the overflows from one facility due to blocking, reneging or balking can become part of the arrival rate to other facilities. The overall performance can be determined by iteratively solving a system of nonlinear equations. The computational method is essentially the same as for the previously studied pure-blocking systems, but now the approach can be used for systems with balking and reneging as well as blocking. We intend to discuss such reduced-load approximations in a subsequent paper.

The BD models are intended to help understand system performance. The BD model simplicity makes it possible to describe performance in detail using an elementary algorithm, but the model requires Markov assumptions such as exponential service-time distributions that may well be seriously violated in practice. The analytical BD model nevertheless can provide important insight. However, to actually predict customer delays in system operation, we contend that it should usually be better not to use the BD model. To predict expected delays and the full delay distribution of arriving customers, we suggest exploiting the actual service-time distributions and the elapsed service times of customers in service. We also suggest exploiting other information enabling the service provider to classify customers. For example, there may be a few known and easily identifiable classes of customers, each with its own service-time distribution. We focus on accurately predicting the delay distribution, but not on precisely what should be told to customers. See Hui and Tse [16], Katz, Larson and Larson [17] and Taylor [26] for discussions of that issue. In some settings (with sophisticated customers), full disclosure may be preferable, i.e., communicating the full delay distribution and possibly other state information.

Given that delays before beginning service can be predicted reasonably well, it is natural to consider not having the customer retry or wait. Instead, the service provider *can call back at a later time.* To do so, the service provider records the calling number when the customer first calls and announces the anticipated future time of the return call, e.g., in about 2 minutes or between 10 and 15 minutes. The accurate delay prediction then helps the service provider accurately predict when the return call can be made.

We now indicate how the rest of this paper is organized. In Section 2 we present what we regard as the traditional BD model to describe performance when some arrivals balk and waiting customers renege after an exponential time. This model can represent both the loss model with rapid retrials and the delay model for the case in which the service provider allows waiting. With retrials, we do not try to directly represent the retrials as in Chapter 7 of Wolff [31]. Instead, assuming that relatively rapid retries are possible, we consider retrying customers to be waiting customers. However, we assume the FCFS service discipline, so that our analysis in Section 2 does not capture the random order of service associated with retrials.

In Section 3 we introduce an alternative BD model to describe the performance when the service provider informs customers about anticipated delays before beginning service or provides state information so that the arriving customers can make this prediction. We relate the state-dependent balking in this setting to the reneging rate in Section 2. The principal change from Section 2 to Section 3 is to replace reneging with balking, but we also allow reneging in Section 3. Thus the model in Section 2 is a special case of the model in Section 3.

In Section 4 we make stochastic comparisons between the models in Sections 2 and 3, showing that state-dependent balking instead of reneging (at comparable rates) leads to fewer customers in the system in steady state. In Section 5 we present some numerical examples giving explicit comparisons. We obtain our numerical results by numerically solving for the performance measures in the BD models. These examples show that the performance in the two scenarios is often remarkably similar. The major difference is that, with balking instead of reneging, customers who do not receive service do not waste time waiting. We also use the numerical examples to show the economies of scale (having fewer groups of larger numbers of servers instead of more groups of smaller numbers of servers).

In Section 6 we present methods for service providers to use to predict the distribution of a customer's delay before beginning service. We present predictions based on the BD models, but also predictions exploiting estimated (non-exponential) service-time distributions and elapsed service times of customers in service. By the classical lack of memory property of the exponential distribution, an elapsed service time does not affect the prediction when the service-time distribution is exponential, but it can have a great impact when the service-time distribution is far from exponential, as when it is a heavy-tail distribution such as the Pareto distribution.

In Sections 7 and 8 we discuss ways to estimate model parameters and validate the BD models. In Section 9 we discuss ways to approximately capture the performance impact of occasional extra long service times. Finally, in Section 10 we briefly discuss other possible deviations from the model assumptions and ways to approximately cope with them. We refer to Boxma and de Waal [4] and Falin [9] for accounts of the literature on queues with reneging and retrials.

2. When Customers Do Not Know the System State

[sec2] In this section we review a reasonably well known birth-and-death (BD) process model for the case in which the system state is not communicated to arriving customers; e.g., see Chapter 2 of Gross and Harris [13] and Chapter 4 of Heyman and Sobel [15]. If a server is not immediately available, then the arriving customer *balks* (leaves immediately) with probability β and waits with probability $1-\beta$. If a server is not immediately available and the customer does not balk, then he *reneges* (abandons later) after an exponential time with mean α^{-1}, if he has not yet begun service. We assume that the system state is not known by customers, so that the parameters α and β cannot depend directly on the number of customers in the system (beyond whether the servers are all busy or not). Once a customer starts service, he stays until service is completed. (It is easy to modify the BD model if this assumption is not reasonable.)

Let the arrival process be a Poisson process with rate λ. Let there be s servers, a waiting room of size r and the first-come first-served service (FCFS) discipline. (The total system capacity is thus $s+r$. An arrival finding $s+r$ customers present is *blocked* (lost). Let the service times be i.i.d. exponential random variables with mean $1/\mu$. Then the birth (arrival) and death (departure) rates are, respectively,

$$\texttt{[eq201]}\lambda_k = \begin{cases} \lambda\ , & 0 \le k \le s-1 \\ \\ \lambda(1-\beta)\ , & s \le k \le s+r-1\ , \end{cases} \tag{2.1}$$

and

$$\texttt{[eq202]}\mu_k = \begin{cases} k\mu\ , & 1 \le k \le s-1 \\ \\ s\mu + (k-s)\alpha\ , & s \le k \le s+r\ . \end{cases} \tag{2.2}$$

When $k > s$, some departures are service completions, while others are abandonments (reneging) and blocked arrivals. The arrival rate λ must be the sum of the rates of service completion, blocking, balking and reneging. We do not discuss how to analyze this BD model here because it is a special case of the model introduced in the next section, which we do analyze.

3. When Customers Know the System State

In this section we consider the case in which customers learn the system state upon arrival. The customers may also receive updates while they are waiting. The customers might be told the number of customers ahead of them in line at all times (e.g., by displays on a monitor with access through a personal computer) and/or they might receive periodic predictions of their remaining time to wait before beginning service (e.g., by telephone announcements with access through a telephone).

Assuming that customers know their preferences, it is natural that customers would respond to this additional information when all servers are busy by replacing reneging after waiting with state-dependent balking; i.e., customers should be able to decide immediately upon arrival whether or not they are willing to join the queue and wait to receive service. Having joined the queue, customers should be much more likely to remain until they begin service. Reneging is even less likely if the customer can see that the remaining time to wait is steadily declining.

Hence in this section we consider an alternative BD model to represent state-dependent balking instead of time-dependent reneging. Since there may still be some reneging in this new situation with additional state information (e.g., because customers change their minds or because progress in the line is slower than anticipated), we also include reneging in the model. However, we are especially interested in the comparison between the model in Section 2 with reneging and the new model in this section in which the reneging is replaced entirely by state-dependent balking. We make a stochastic comparison in Section 4.

As before, there is a Poisson arrival process with rate λ and s servers, each with exponential service times having mean μ^{-1}. There is a waiting room of size r and the FCFS service discipline. An arrival encountering a full system is blocked. Paralleling Section 2, the customers are assumed to be willing to wait until starting service a random time that is exponentially distributed with mean $1/\alpha$. These times for different customers are assumed to be mutually independent. However, now the customer learns the system state upon arrival and decides whether or not to balk. If the number seen by the arrival (not including the arrival) is less than or equal to $s-1$, then the new arrival enters service immediately. If the number seen by the arrival is $s+k$ for $0 \le k \le r-1$, then the arrival may elect to balk (leave immediately) or join the queue. Paralleling Section 2, each customer finding all servers busy balks with probability β. However, the customer may also elect to balk depending on the system state. We stipulate that the customer joins with the probability that a server becomes free before he would abandon. Let S_k be the time required from arrival until a server first becomes available for this customer, as a function of k, assuming that departures occur only by service completions (not considering reneging by customers in queue ahead of the current customer), and let T be the time that this customer would have reneged in Section 2. (We assume that the actual service times are not known.) Then the arrival finding $s+k$ customers in the system upon arrival (not counting himself) joins with probability

$$\texttt{[eq301]}\, q_k \equiv P(T > S_k) , \quad 0 \le k \le r-1 . \tag{3.1}$$

Since S_k has the distribution of the sum of $k+1$ exponentials each with mean $1/s\mu$ and T has an exponential distribution with mean $1/\alpha$, we can exploit Laplace transforms to calculate q_k explicitly. In particular,

$$\texttt{[eq302]}\, q_k = \int_0^\infty e^{-\alpha t} P(S_k = dt) = \left(\frac{s\mu}{s\mu+\alpha}\right)^{k+1} . \tag{3.2}$$

We also indicate several alternatives to (3.1) and (3.2). The first alternative is intended to represent the case in which the service provider communicates the expected delay when there are $s+k$ customers in the system. Then we would replace S_k in (3.1) by its mean, i.e., we would use

$$\texttt{[eq303]}\, \bar{q}_k \equiv P(T > ES_k) = e^{-\alpha(k+1)/s\mu} , \quad k \ge 0 . \tag{3.3}$$

Note that when k is large, S_k will tend to be relatively close to ES_k by the law of large numbers. Directly, we can see that, if k and s are suitably large, then (3.2) will be close to (3.3), i.e.,

$$\begin{aligned} \left(\frac{s\mu}{s\mu+\alpha}\right)^{k+1} &= \left(1 - \frac{\alpha}{s\mu+\alpha}\right)^{k+1} \\ \texttt{[eq303a]} &= \left(1 - \frac{(k+1)\alpha}{(k+1)(s\mu+\alpha)}\right)^{k+1} \approx e^{-(k+1)\alpha/(s\mu+\alpha)} \approx e^{-(k+1)\alpha/s\mu} . \end{aligned} \tag{3.4}$$

In general, $\bar{q}_k \geq q_k$.

The analysis leading to (3.2) and (3.3) suggests that the probability a customer joins the queue (does not balk) when he finds $s+k$ in system should be of the general form $\xi\zeta^{-k}$ for parameters ξ and η with $0 \leq \xi \leq 1$ and $0 \leq \eta \leq 1$. In practice the balking probability as a function of k needs to be estimated. This blocking probability should depend on the information supplied to the customer.

We now define a BD model representing state-dependent balking. Since there may still be some reneging, we include state-dependent reneging as well. The birth (arrival) and death (departure) rates are, respectively,

$$\texttt{[eq304]}\lambda_k = \begin{cases} \lambda\ , & 0 \leq k \leq s-1 \\ \\ \lambda(1-\beta)q_{k-s}\ , & s \leq k \leq s+r-1 \end{cases} \tag{3.5}$$

and

$$\texttt{[eq305]}\mu_k = \begin{cases} k\mu\ , & 1 \leq k \leq s-1 \\ \\ s\mu + (k-s)\delta_{k-s}\ , & s \leq k \leq s+r\ . \end{cases} \tag{3.6}$$

In (3.6) we have allowed general state-dependent reneging rate for each waiting customer, δ_k, but we will usually consider the special case in which $\delta_k = \delta$. The model in Section 2 corresponds to that special case with the parameter pair (α, δ) here set equal to $(0, \alpha)$.

We now indicate how to numerically solve for the steady-state probabilities p_k. Since the larger probabilities should be near s (assuming that s is reasonably well chosen), it is convenient to solve for the steady-state distribution recursively starting at s. Let $x_s = 1$,

$$\texttt{[eq203]}x_{s+k+1} = \frac{\lambda_{s+k}x_{s+k}}{\mu_{s+k+1}} = \frac{\lambda(1-\beta)q_k x_{s+k}}{s\mu + (k+1)\delta_{k+1}}\ , \quad 0 \leq k \leq r-1\ , \tag{3.7}$$

and

$$\texttt{[eq204]}x_{k-1} = \frac{\mu_k x_k}{\lambda_{k-1}} = \frac{k\mu x_k}{\lambda}\ , \quad 1 \leq k \leq s\ . \tag{3.8}$$

Then, let

$$\texttt{[eq205]}y = \sum_{k=0}^{s+r} x_k \tag{3.9}$$

and

$$\texttt{[eq206]}p_k = x_k/y\ , \quad 0 \leq k \leq s+r\ . \tag{3.10}$$

So far the results have been quite standard, but now we go on to compute the probability of completing service and the mean, variance and full distribution of the conditional response time (time to complete service) given that service is completed. We also compute the probability that a customer reneges and the mean, variance and full distribution of the conditional time to renege given that the customer reneges.

Since the arrival process is Poisson, the state seen by arrivals is the same as at an arbitrary time by the Poisson-Arrivals-See-Time-Average (PASTA) property; see Section 5.16 of Wolff [31]. Let γ_k be the probability that the k^{th} customer in line abandons in the next departure event (assuming each customer is equally likely to abandon) and let m_k be the mean time to the next departure event, in both cases considering only the first $s+k$ customers in the system; i.e.,

$$\texttt{[eqQ2]}\gamma_k = \frac{\delta_k}{s\mu + k\delta_k} \quad \text{and} \quad m_k = \frac{1}{s\mu + k\delta_k} \ . \tag{3.11}$$

Then the probability that customer $s+k$ eventually receives service is

$$\texttt{[eqQ3]}\Gamma_k = (1-\gamma_k)(1-\gamma_{k-1})\dots(1-\gamma_1) \tag{3.12}$$

for γ_k in (3.11). Then the probability that a new arrival eventually completes service, is

$$\texttt{[eq306]}P(S) = \left(\sum_{k=0}^{s-1} p_k\right) + \sum_{k=0}^{r-1} p_{s+k}(1-\beta)q_k\Gamma_{k+1} \ . \tag{3.13}$$

Let C be the response time. (We let C be 0 when service is not completed.) Then, using properties of the exponential distribution, we obtain

$$\texttt{[eq307]}EC = \left(\sum_{k=0}^{s-1} p_k\right)\frac{1}{\mu} + \sum_{k=0}^{r-1} p_{s+k}(1-\beta)q_k\Gamma_{k+1}\left(\frac{1}{\mu} + \sum_{j=1}^{k+1} m_j\right) \tag{3.14}$$

and

$$\texttt{[eq308]}EC^2 = \left(\sum_{k=0}^{s-1} p_k\right)\frac{2}{\mu^2} + \sum_{k=0}^{r-1} p_{s+k}(1-\beta)q_k\Gamma_{k+1}(V_{k+1} + M_{k+1}^2) \tag{3.15}$$

where

$$\texttt{[eqQ4]}V_{k+1} = \frac{1}{\mu^2} \ + \ \sum_{j=1}^{k+1} m_j^2 \tag{3.16}$$

and

$$\texttt{[eqQ5]}M_{k+1} = \frac{1}{\mu} \ + \ \sum_{j=1}^{k+1} m_j \ . \tag{3.17}$$

Then the first and second moments of the conditional time to complete service given that service is completed are

$$\texttt{[eq309]}E(C|S) = EC/P(S) \quad \text{and} \quad E(C^2|S) = EC^2/P(S) \ . \tag{3.18}$$

The conditional variance and standard deviation are then

$$\texttt{[eq311]}Var(C|S) = E(C^2|S) - (E(C|S))^2 \tag{3.19}$$

and

$$[\texttt{eq312}] SD(C|S) = \sqrt{Var(C|S)}\ . \tag{3.20}$$

Now let $\hat{c}(s) \equiv Ee^{-sC}$ be the Laplace transform of C. Paralleling (3.14), we have

$$[\texttt{eqE1}] \hat{c}(s) = \left(\sum_{k=0}^{s-1} p_k\right)\left(\frac{\mu}{\mu+s}\right) + \sum_{k=0}^{r-1} p_{s+k}(1-\beta)q_k\Gamma_{k+1}\hat{d}_{k+1}(s)\ , \tag{3.21}$$

where

$$[\texttt{eqE2}] \hat{d}_{k+1}(s) = \left(\frac{\mu}{\mu+s}\right)\Pi_{j=1}^{k+1}\left(\frac{m_j}{m_j+s}\right)\ . \tag{3.22}$$

We can now easily calculate $P(X > t)$ for any desired t by numerically inverting its Laplace transform $(1-\hat{c}(s))/s$, e.g., by using the Fourier-series method described in Abate and Whitt [1]. The associated conditional response-time distribution is

$$[\texttt{eqE3}] P(C > t|S) = P(C > t)/P(S)\ . \tag{3.23}$$

Let R be the event that an arrival eventually reneges and let A be the time to renege. Then, by essentially the same reasoning,

$$[\texttt{eqQ6}] P(R) = \sum_{k=0}^{r-1} p_{s+k}(1-\beta)q_k(1-\Gamma_{k+1})\ , \tag{3.24}$$

$$[\texttt{eqQ7}] EA = \sum_{k=1}^{r} p_{s+k-1}(1-\beta)q_{k-1}EA(k) \tag{3.25}$$

and

$$[\texttt{eqQ8}] EA^2 = \sum_{k=1}^{r} p_{s+k-1}(1-\beta)q_{k-1}EA(k)^2\ , \tag{3.26}$$

where

$$\begin{aligned} EA(k) &= \gamma_k m_k + (1-\gamma_k)\gamma_{k-1}(m_k+m_{k-1}) + (1-\gamma_k)(1-\gamma_{k-1})\gamma_{k-2}(m_k+m_{k-1}+m_{k-2}) \\ [\texttt{eq216}] &\quad + \ldots + (1-\gamma_k)\ldots(1-\gamma_2)\gamma_1(m_k+\ldots+m_1) \end{aligned} \tag{3.27}$$

and

$$\begin{aligned} EA(k)^2 &= \gamma_k 2m_k^2 + (1-\gamma_k)\gamma_{k-1}(m_k^2+m_{k-1}^2+(m_k+m_{k-1})^2) \\ [\texttt{eq217}] &\quad + \ldots + (1-\gamma_k)(1-\gamma_{k-1})\ldots(1-\gamma_2)\gamma_1(m_k^2\ldots+m_1^2+(m_k+\ldots+m_1)^2) \end{aligned} \tag{3.28}$$

The associated conditional moments are

$$[\texttt{eq220}] E(A|R) = EA/P(R) \quad \text{and} \quad E(A^2|R) = EA^2/P(R)\ , \tag{3.29}$$

for $P(R)$ in (3.24). Finally, the conditional variance and standard deviation are

$$\texttt{[eq222]}Var(A|R) = E(A^2|R) - (E(A|R))^2 \tag{3.30}$$

and

$$\texttt{[eq223]}SD(A|R) = \sqrt{Var(A|R)}\ . \tag{3.31}$$

Now let $\hat{a}(s) \equiv e^{-sA}$ be the Laplace transform of A. Paralleling (3.25), we have

$$\texttt{[eqE4]}\hat{a}(s) = \sum_{k=0}^{r-1} p_{s+k}(1-\beta)q_k(1-\Gamma_{k+1})\hat{a}_k(s)\ , \tag{3.32}$$

where

$$\texttt{[eqE5]}\hat{a}_k(s) = \left(\frac{m_k}{m_k+s}\right)\sum_{j=0}^{k-1}\gamma_{k-j}\Pi_{\ell=1}^{j}\left[(1-\gamma_{k-\ell+1})\left(\frac{m_{k-\ell}}{m_{k-\ell}+s}\right)\right]\ . \tag{3.33}$$

Paralleling $P(C>t)$ above, we can compute $P(A>t)$ by numerically inverting its Laplace transform $(1-\hat{a}(s))/s$. Then the conditional distribution of the time to renege given reneging is

$$\texttt{[eqE6]}P(A>t|R) = P(A>t)/P(R)\ . \tag{3.34}$$

Finally, the probability of blocking is p_{s+r}, so that the probability of balking is

$$\texttt{[eqQ9]}P(\text{balking}) = 1 - P(S) - P(R) - p_{s+r}\ . \tag{3.35}$$

4. Stochastic Comparisons

[sec4] The consequences of informing customers about anticipated delays are not entirely clear. It seems that customers should prefer this additional information and that the greatest benefit will stem from improved customer satisfaction. However, the impact on congestion is less clear. First, the rate of customer service and the steady-state number of customers in the system might both increase because the fixed balking rate β might decrease and the arrival rate λ might increase. It thus might be necessary to increase the number of servers s, i.e., better service might mean more business.

In this section we make comparisons assuming that the parameters remain unchanged. Intuitively, it seems that balking upon arrival instead of joining the queue and later reneging should lead to fewer customers in the system, provided that the chance of balking relates appropriately to the chance of reneging, as in the construction in Section 3, in particular, assuming (3.2). We now show that a strong comparison is possible. In particular, we establish *likelihood ratio* (MLR) ordering. See Chapter 1 of Shaked and Shanthikumar [24] for background on stochastic orderings.

Consider two random variables X_1 and X_2 with values in the state space $\{0, 1, \ldots, s\}$, $1 \le s \le \infty$, that have probability mass functions (pmf's) that are positive for all states. We say that X_1 is less than or equal to X_2 in the *likelihood ratio* (LR) ordering and write $X_1 \le_{lr} \le X_2$ if

$$\texttt{[eq401]}\,\frac{P(X_1 = k+1)}{P(X_1 = k)} \le \frac{P(X_2 = k+1)}{P(X_2 = k)}\ , \quad 0 \le k \le s-1\ . \tag{4.1}$$

We say that X_1 is *stochastically less than or equal to* X_2 and write $X_1 \le_{st} X_2$ if

$$\texttt{[eq402]}\,P(X_1 \ge k) \le P(X_2 \ge k)\ , \quad 0 \le k \le s\ . \tag{4.2}$$

The LR order implies stochastic order. Indeed, the LR order is equivalent to stochastic order holding under conditioning for all intervals; i.e., $X_1 \le_{lr} X_2$ if and only if

$$\texttt{[eq403]}\,(X_1 | a \le X_1 \le b) \le_{st} (X_2 | a \le X_2 \le b) \tag{4.3}$$

for all a and b with $a < b$; see p. 29 of Shaked and Shanthikumar.

We now present a sufficient condition for the steady-state distributions of BD processes to be ordered in the LR ordering. This result is a special case of Theorem 5 of Smith and Whitt [25] (which applies to more general processes).

Theorem 4.1 `[thm401]` *Consider two BD processes with common state space* $\{0, 1, \ldots, s\}$, *birth rates* $\lambda_k^{(i)}$, *death rates* $\mu_k^{(i)}$ *and steady-state random variables* N_i, $i = 1, 2$. *If*

$$\texttt{[eq404]}\,\frac{\lambda_k^{(1)}}{\mu_{k+1}^{(1)}} \ge \frac{\lambda_k^{(2)}}{\mu_{k+1}^{(2)}} \quad \textit{for } 0 \le k \le s-1\ , \tag{4.4}$$

then

$$N_1 \ge_{lr} N_2\ .$$

We now compare the processes in Sections 2 and 3, where the model in Section 3 has no reneging.

Theorem 4.2 `[thm402]` *Consider the BD processes introduced in Sections 2 and 3, using (3.2), with common parameters* $\lambda, \mu, \alpha, \beta, s$ *and* r *and no reneging for the model in Section 3, i.e., with* $\delta_k = 0$. *Let the model with reneging in Section 2 be indexed by superscript 1 and the other model by superscript 2. Let* $\lambda_k^{(i)}, \mu_k^{(i)}$ *and* N_i *denote the birth rates, death rates and steady-state number of customers present in model* i. *Then*

$$N_1 \ge_{lr} N_2\ .$$

Proof. By Theorem 4.1, it suffices to establish (4.4). For $s + k \geq 0$, $\lambda_k^{(1)} = \lambda_k^{(2)} = \lambda$ and $\mu_{k+1}^{(1)} = \mu_{k+1}^{(2)} = (k+1)\mu$. For $k \geq 0$,

$$\frac{\lambda_{s+k}^{(1)}}{\mu_{s+k+1}^{(1)}} = \frac{\lambda(1-\beta)}{s\mu + (k+1)\alpha}\ , \quad \frac{\lambda_{s+k}^{(2)}}{\mu_{s+k+1}^{(2)}} = \frac{\lambda(1-\beta)q_k}{s\mu}\ ,$$

so that it suffices to show that

$$\texttt{[eq405]}\,\frac{\lambda_{s+k}^{(2)}}{\lambda_{s+k}^{(1)}} = q_k \equiv \left(\frac{s\mu}{s\mu+\alpha}\right)^{k+1} \leq \frac{s\mu}{s\mu + (k+1)\alpha} = \frac{\mu_{s+k+1}^{(2)}}{\mu_{s+k+1}^{(1)}} \tag{4.5}$$

for $0 \leq k \leq r-1$. However, (4.5) holds because, by the binomial theorem, $(1+x)^k \geq 1 + kx$ for all $x > 0$ and all positive integers k. ∎

The stochastic comparison we have made between the two modes of operation in Theorem 4.2 assumes that the basic parameter tuple $(\lambda, \mu, \alpha, \beta, s, r)$ is the same for both systems. However, if we change the way the system operates, then these parameters may change too, leading to more complex comparisons. We can describe how each system separately responds to changes in the parameters, though. For simplicity, let $\delta_k = \delta$ in Section 3, then the model there depends on the parameter tuple $(\lambda, \mu, \alpha, \beta, \delta, s, r)$.

Theorem 4.3 `[thm403]` *Consider one of the systems in Section 2 or 3. Let N_i be the steady-state number of customers in system i with parameter tuple $(\lambda^{(i)}, \mu^{(i)}, \alpha^{(i)}, \beta^{(i)}, \delta^{(i)}, s, r)$, $i = 1, 2$. If $\lambda^{(1)} \leq \lambda^{(2)}$, $\mu^{(1)} \geq \mu^{(2)}, \alpha^{(1)} \geq \alpha^{(2)}$, $\beta^{(1)} \geq \beta^{(2)}$ and $\delta^{(i)} \geq \delta^{(2)}$, then then*

$$N_1 \leq_{lr} N_2\ .$$

Proof. It is easy to see that $\lambda_k^{(1)} \leq \lambda_k^{(2)}$ and $\mu_{k+1}^{(1)} \geq \mu_{k+1}^{(2)}$ for all k, $0 \leq k \leq s + r - 1$, so that (4.4) holds. Hence, we can apply Theorem 4.1. ∎

From Theorem 4.3 it is not evident how the long-run balking and reneging rates respond to increases in the parameters α, β and δ. It is intuitively clear that the long-run rates should increase, but if we increase β, then the steady-state distribution decreases, so that there is less opportunity for balking. Nevertheless, we can establish the desired comparison by exploiting a sample-path comparison.

Theorem 4.4 `[thm404]` *Consider one of the systems in Section 2 or 3.*

(a) If α increases, then the long-run reneging rate (Section 2) or balking rate (Section 3) increases.

(b) If β increases, then the long-run balking rate increases.

(c) If α, β and δ increase, then the long-run rate of service completions decreases.

Proof. We only consider part (a) for the system in Section 2, because the reasoning is the same in the other cases. As in Whitt [27], it is possible to construct the two systems on the same sample space so that the sample paths are ordered (a coupling). Let the two systems be indexed by i, where $\alpha^{(1)} < \alpha^{(2)}$. Let $N_i(t)$ be the number of customers in system i as a function of time. Let the two systems both start out empty. We can generate all events from a common Poisson process with a constant rate $\gamma \equiv \lambda + s\mu + r\alpha$. Then we determine the nature of the events according to the birth and death rates. For example, with probability λ/γ, the event is an external arrival. If the state is $k < s$, then with probability $k\mu/\gamma$, the event is a service completion, while with probability $(\gamma - \lambda - k\mu)/\gamma$ the event is a fictitious event, leading to no state change. Whenever the two sample paths coincide with $s + k$ customers present for $k \geq 1$, let service completions be the same in both systems and let there be reneging in the system with parameter $\alpha^{(2)}$, where $\alpha^{(2)} > \alpha^{(1)}$, whenever there is reneging in the system with parameter $\alpha^{(1)}$. However, there may be additional reneging in system 2, making $N^{(2)}(t) \leq N^{(1)}(t)$. Whenever $N^{(2)}(t) \leq N^{(1)}(t)$, the service completion rates and are greater for system 1. Hence, let there be a service completion in system 1 whenever there is one in system 2. This allows extra service completions in system 1. Also, let there be a balking event in system 1 whenever there is one in system 2, which allows extra balking events in system 1. With this construction, a gap $N^{(1)}(t) - N^{(2)}(t)$ can only be created and grow by excess reneging in system 2. This gap may be reduced in several ways, including by subsequent reneging in system 1, but the cumulative number of customers reneging always stays ahead for system 2. ∎

Remark. For the model in Section 2, the proof of Theorem 4.4 shows that the long-run service completion and balking rates both decrease when α increases. When α and β both increase, we can deduce that the long-run service completion rate decreases, but not how the long-run reneging and balking rates are affected.

5. Numerical Examples

[sec5] We now illustrate how the BD models can be used by considering a few numerical examples. In Theorem 4.2, we established an ordering between the two systems with common parameter tuples $(\lambda, \mu, \alpha, \beta, s, r)$. However in numerical examples we have found that in many respects the two systems with common parameter tuples behave very similarly. The main difference is that, for

the system in Section 2, some customers who do not eventually receive service spend time waiting before reneging. This wasted customer effort is eliminated by predicting delays, if the prediction leads to the model in Section 3. Throughout this section we use definition (3.2).

Example 5.1. Economies of Scale

In addition to comparing the two systems with common parameter tuples, our first example illustrates the economies of scale. In particular, we consider both systems (with and without reneging, as in Sections 2 and 3) with $s = 4 \times 10^k$ for $k = 0, 1, 2$ and 3. In each case, we let $\lambda = s$, $\mu = 1.0, \alpha = 1.0$ and $\beta = 0.2$. We choose r to be sufficiently large so that blocking is negligible. With this parameter choice, the system with $s = 4 \times 10^k$ corresponds to the combination of 10 identical systems with $s = 4 \times 10^{k-1}$. We have resource sharing in the sense of Smith and Whitt [25].

Numerical results for these cases are presented in Table 1. Since $(x)^+ = \max\{x, 0\}$, $E(N-s)^+$ there is the expected number of customers waiting. Table 1 shows that the two systems do not differ much, with the difference decreasing as s increases. In all cases, the probability that an arrival is eventually served are very close for the two systems. Table 1 also shows that all measures of performance improve as s increases, thus quantifying the economies of scale.

It is interesting to contrast the balking and reneging examples with the pure-loss model, which otherwise has the same parameters. The probability of eventually being served in the associated M/M/s/0 loss model is 0.639, 0.884, 0.961 and 0.9875 for $s = 4 \times 10^k$ and $k = 0, 1, 2$ and 3. The difference is substantial for smaller s, but negligible for larger s. For larger s, the balking acts like blocking. When $\lambda = 4000$ and $\beta = 0.2$, the arrival rate drops to 3200 when all servers are busy. In that case, s acts much like an upper barrier. Indeed, in that case, the conditional mean queue length given all servers are busy is only $E(N-s)^+/P(N \geq s) = 3.95$. ∎

performance measures	$s = 4$		$s = 40$	
	reneging	only balking	reneging	only balking
$P(N \geq s)$	0.501	0.493	0.335	0.333
$E(N-s)^+$	0.498	0.445	0.816	0.796
EN	3.60	3.53	37.3	37.3
$SD(N)$	1.74	1.67	4.86	4.83
$P(renege)$	0.124	0	0.020	0
$P(served)$	0.775	0.772	0.913	0.912
$E(C\|S)$	1.115	1.144	1.021	1.022
$SD(C\|S)$	1.026	1.046	1.001	1.001
$E(A\|R)$	0.282	--	0.069	--
	$s = 400$		$s = 4000$	
	reneging	only balking	reneging	only balking
$P(N \geq s)$	0.162	0.162	0.0593	0.0593
$E(N-s)^+$	0.589	0.589	0.234	0.234
EN	387.0	387.0	3953.	3953.
$SD(N)$	13.1	13.1	38.9	38.9
$P(renege)$	0.0015	0	0.00006	0
$P(served)$	0.966	0.966	0.9881	0.9881
$E(C\|S)$	1.0015	1.0015	1.0000	1.0000
$SD(C\|S)$	1.0000	1.0000	1.0000	1.0000
$E(A\|R)$	0.110	--	0.0012	--

Table 1. A comparison of the two service schemes as a function of system size, $s = 4 \times 10^k$ for $k = 1, 2, 3$ and 4. In all cases $\lambda = s$, $\mu = \alpha = 1$ and $\beta = 0.2$. The variable N is the steady-state number of customers in the system, C is the time to complete service and A is the time to abandon.

Example 5.2. Heavy Loads

The systems in Sections 2 and 3 do not differ when all servers are not busy. Thus, the difference should increase as the load increases. We next illustrate the larger differences that are possible with higher loads. For this example, we let $s = 10$, $\mu = 1.0$, $\alpha = 1.0$ and $r = 50$. We consider two cases: In the first case, we let $\lambda = 20$ and $\beta = 0.2$; in the second case we let $\lambda = 40$ and $\beta = 0.5$. With the conventional definition of traffic intensity $\rho \equiv \lambda/s\mu$, $\rho = 2.0$ and 4.0 in the two cases. Numerical results for these two cases are displayed in Table 2.

In Table 2 the differences between the two systems are greater than in Table 1, but still not large. The probability of being eventually served and the mean and standard deviation of the conditional time to be served are very close. The greatest differences are in EN and $SD(N)$, the mean and standard deviation of the steady-state number of customers in the system.

performance measures	$\lambda = 20,\ \mu = 1.0,\ \alpha = 1.0,$ $\beta = 0.2,\ s = 10,\ r = 50$		$\lambda = 40,\ \mu = 1.0,\ \alpha = 1.0,$ $\beta = 0.5,\ s = 10,\ r = 50$	
	reneging	only balking	reneging	only balking
$P(N \geq s)$	0.970	0.958	0.9981	0.9955
$E(N-s)^+$	6.17	4.66	10.04	6.84
EN	16.1	14.6	20.0	16.8
$SD(N)$	3.90	3.09	4.43	3.16
$P(renege)$	0.308	0	0.251	0
$P(served)$	0.498	0.497	0.250	0.250
$E(C\|S)$	1.44	1.47	1.65	1.68
$SD(C\|S)$	1.04	1.065	1.045	1.080
$E(A\|R)$	0.293	--	0.356	--

Table 2. A comparison of the two service schemes in Sections 2 and 3 under heavy loadings.

Note that the most serious detrimental effect of the heavy loads is the low proportion of customers served. The delays experienced by those customers served are not especially large. These results show that a focus on the delays experienced by served customers, while ignoring the customers lost to balking or reneging, can seriously overestimate the quality of service provided.

Also note that the performance is quite different from the M/M/s/r model without balking or reneging. Then the steady-state number N is close to $s+r$, which in Table 2 would be 60. The probability of being served is about the same, however. In the setting of Table 2 the blocking is negligible. The high blocking in M/M/s/r is replaced by balking and reneging in these cases. ∎

At first glance, it might be thought that in the setting of Section 2 the reneging rate α might be reasonably well estimated by the reciprocal $E(A|R)$, the expected time to renege given that reneging occurs. However, it can be much less. Note that we consistently have

$$\texttt{[eqR1]}\, E(A|R) \leq 1/\alpha \ . \tag{5.1}$$

This must occur because the sequence of reneging times is censored. (Many customers are served before they have a chance to renege.)

6. Predicting Future Delays

`[sec6]` In order for the service provider to accurately predict delays before arrivals can begin service, the service provider needs to be able to accurately estimate future delays given the system state. Given the model in Section 3, it is relatively easy to accurately estimate delays. Since there should be negligible reneging, the waiting time before starting service for an arrival finding $s+k$ customers in the system is the sum of $k+1$ i.i.d. exponential random variables each with mean $1/s\mu$. Hence, the mean and standard deviation of the steady-state waiting time before starting service, W are

$$\texttt{[eq601]} EW = \frac{k+1}{s\mu} \quad \text{and} \quad SD(W) = \frac{\sqrt{k+1}}{s\mu} \ . \tag{6.1}$$

It is reasonable to normalize by the mean service time, so that waiting times are viewed in relation to mean service times. This is equivalent to setting $\mu = 1$. Then

$$\texttt{[eq602]} EW = \frac{k+1}{s} \quad \text{and} \quad SD(W) = \frac{\sqrt{k+1}}{s} \ . \tag{6.2}$$

The formulas in (6.2) show the advantage of large scale. When s is big, either EW and $SD(W)$ are both small (when k is small) or the ratio $SD(W)/EW$ is small (when k is large). Moreover, when k is not too small, we can apply the central limit theorem to deduce that W is approximately normally distributed with the mean and standard deviation just determined.

We have suggested that the model in Section 2 should be replaced by the model in Section 3 with negligible reneging when the service provider predicts delays. However, it is also possible to predict delays when there is significant reneging. In the setting of Section 2, when a customer finds $s+k$ customers in the system, the delay can again be represented as the sum of $k+1$ independent exponential random variables, but now they are not identically distributed. The mean and standard deviation become

$$\texttt{[eq603]} EW = \sum_{j=0}^{k} \frac{1}{s\mu + j\alpha} \quad \text{and} \quad SD(W) = \left[\sum_{j=0}^{k} \frac{1}{(s\mu + j\alpha)^2} \right]^{1/2} \ . \tag{6.3}$$

The modification (6.3) can be important if there is some reneging, even though the service provider predicts delays. If we ignore reneging, then the delay predictions will be somewhat pessimistic.

However, the accuracy of the delay prediction above depends strongly on the exponential service-time assumption. For other service-time distributions, the remaining service time depends on the elapsed service time. Thus, for non-exponential service-time distributions, we can more accurately predict the delay of a new arrival if we exploit the elapsed service times (ages) of the customers in service. Moreover, it may be possible to classify customers into different types, where each type has a very different service-time distribution. This classification may be done before or after service has begun.

Henceforth, assume that the classification has been done before service begins, so that customer i before starting service has service-time cdf G_i. The actual service-time cdf G_i should be easily estimated directly from the observed service times, assuming that there is no reneging after service has begun and that service times in progress are not altered by system state. In practice, this last possibility should be checked. It can be checked by estimating service-time distributions conditional on the number in system when service starts. With significant reneging, the estimation procedures should account for censoring.

Let $G_i(t|x)$ be the cdf of the conditional remaining service time, conditional on an elapsed service time (age) of x. Clearly,

$$\texttt{[eq604]}\, G_i(t|x) = \frac{G_i(t+x)}{1-G_i(x)} \ , \quad t \geq 0 \ . \tag{6.4}$$

Suppose that the service provider keeps tract of the starting time for each service in process, so that at the time of a new arrival, the ages of the service times of all customers in service are known. If additional prediction is done after service has started (using service time), then $G_i(t|x_i)$ could be estimated directly instead of by (6.4).

At this point, one approach is to use an infinite-server approximation, as in Duffield and Whitt [8]. Let $D(t)$ be the number of departures by time t. With an infinite-server approximation, we optimistically act as if all customers in the system are in service. This leads to the approximation

$$\texttt{[eq612]}\, ED(t) \approx \sum_{i=1}^{s} G_i(t|x_i) + \sum_{i=s+1}^{s+k} G_i(t) \tag{6.5}$$

Similarly, we can approximate the variance by

$$\texttt{[eqB1]}\, VarD(t) \approx \sum_{i=1}^{s} G_i(t|x_i)(1-G_i(t|x_i)) + \sum_{i=s+1}^{s+k} G_i(t)(1-G_i(t)) \ , \tag{6.6}$$

assuming that these are independent (non-identically distributed) trials. Then we estimate the mean waiting time before the new arrival can start service as

$$\texttt{[eq609]}\, EW \approx \min\{t > 0 : ED(t) = k+1\} \ . \tag{6.7}$$

We emphasize that (6.7) is an approximation. The actual waiting time is

$$\texttt{[eqA2]}\, W = \min\{t \geq 0 : D(t) = k+1\} \ , \tag{6.8}$$

We cannot actually obtain the mean of W in (6.8) by replacing $D(t)$ in (6.8) by its mean, but this is the candidate approximation proposed by Duffield and Whitt [8].

However, it may be somewhat too optimistic to act as if waiting customers start service immediately. Hence, we now introduce a refinement in which waiting customers are allowed to start service in the future. For this purpose, let $D_s(t)$ be the number of the current s customers in service that will have departed t time units later. Then, given the s ages $x_1, \ldots, x_s$, its expected value is

$$\texttt{[eq605]} ED_s(t) = \sum_{i=1}^{s} G_i(t|x_i) \ . \tag{6.9}$$

To estimate when the waiting customers start service, let

$$\texttt{[eq606]} t_j = \min\{t \geq 0 : ED_s(t) = j\} \ . \tag{6.10}$$

Note that, like (6.7), (6.10) is an approximation because it is a first passage time for the mean instead of the first passage time for the process itself. Equation (6.10) is exact in the case of deterministic (possibly different) service times, though.

Again let $D(t)$ be the total number of departures by time t. Then we can estimate its mean by

$$\texttt{[eq607]} ED(t) \approx \sum_{i=1}^{s} G(t|x_i) + \sum_{j=s+1}^{s+k} G_j(t - t_{j-s}) \tag{6.11}$$

for t_j in (6.10). Paralleling (6.6), we estimate the variance of $D(t)$ by

$$\texttt{[eq608]} VarD(t) \approx \sum_{i=1}^{s} G_i(t|x_i)(1 - G_i(t|x_i) + \sum_{j=s+1}^{s+k} G_j(t - t_{j-s})(1 - G_j(t - t_{j-s})) \ . \tag{6.12}$$

Note that (6.11) and (6.12) do not account for departures from waiting customers after the first k before some of the accounted for departures, but this discrepancy should be relatively small if k is not large compared to s.

Given (6.10) and (6.11), we can estimate the full waiting-time distribution (approximately). We start by the observation that $D(t)$, being a sum of independent random variables, should be approximately normally distributed, by virtue of the central limit theorem for non-identically distributed random variables; see p. 262 of Feller [10]. Let $N(0,1)$ denote a standard (mean 0, variance 1) normal random variable and let Φ be its cdf.

Let

$$\texttt{[eqA1]} w_x = \min\{t \geq 0 : ED(t) + xSD(D(t)) = k + 1\} \ , \tag{6.13}$$

where SD is the standard deviation. Since the random waiting time W is defined by (6.8), it is natural to use the approximation

$$\texttt{[eqA3]} P(W \geq w_x) \approx P(D(t) \leq ED(t) + xSD(D(t)) \approx P(N(0,1) \leq x) \approx \Phi(x) \ . \tag{6.14}$$

Now, for $0 < \alpha < 1$, let $x_\alpha = \Phi^{-1}(\alpha)$, i.e.; choose x_α so that $\Phi(x_\alpha) = \alpha$. Then w_{x_α} is the approximate $(1 - \alpha)$-percentile of the distribution of W, i.e.,

$$\texttt{[eqA4]} P(W \geq w_{x_\alpha}) = \alpha \ . \tag{6.15}$$

From (6.15), we can obtain the complementary cdf $P(W > w)$ and then compute any desired summary characteristic. For example, the mean is

$$\texttt{[eqA5]}EW = \int_0^\infty P(W > w)dw \ . \tag{6.16}$$

We also suggest using the median, either directly or as an estimate of the mean, which leads to (6.7) with $ED(t)$ in (6.11) instead of (6.5). If we want to be conservative, then we can include the standard deviation in the mean waiting-time approximation, e.g., by replacing (6.7) with

$$\texttt{[eq610]}EW \approx \min\{t \geq 0 : E(D(t)) - cSD(D(t)) = k + 1\} \tag{6.17}$$

for some constant c, e.g., $c = 1$. From above, we see that (6.17) is tantamount to using the estimate of the $(1 - \alpha)$-percentile of the distribution for $x_\alpha = -c$.

We calculate (6.7), (6.13) or (6.17) by first calculating (6.9) for a set $\mathcal{T}$ of time points t, using (6.4). We then approximate the times t_j in (6.10) from among the time points in $\mathcal{T}$ considered in (6.9). Then, given the times t_j in (6.10), we compute the second terms of (6.11) and (6.12) for all times t in $\mathcal{T}$. We approximate EW in (6.7), (6.13) and (6.17) by again only considering the time points in $\mathcal{T}$. Assuming that s is relatively large, the approximation (6.7) can be justified by the law of large numbers, as in Duffield and Whitt [8].

Example 6.1. Validation of the Delay Prediction

We can quickly validate the main approximation above by making comparison to exact results for the BD model. For example, suppose that $s = 100$ and $\mu = 1$. Let the initial number in the system be 130. By (6.2), the waiting time before a new arrival can begin service has mean 0.31, and standard deviation 0.056. In contrast, the infinite-server approximation for the mean is that value of t for which $130e^{-t} = 99$ or $t = -\log(99/130) = 0.27$ which is optimistic, as indicated before. The refined approximation in (6.10) has start times $t_j = -\log(1 - (j/100))$, $1 \leq j \leq 30$. Then, using (6.11) and (6.7), we get the approximation $EW \approx 0.313$. However, the big advantages of the approximation procedure are with non-identical or non-exponential service-time distributions. The approximation in such cases can be validated by simulation. ∎

In some applications we may be especially interested in the waiting times of the first few customers in queue. In that case, it is feasible to calculate the exact waiting-time distribution. Let W_k be the waiting time of the k^{th} customer in line. Then the complementary cdf of the waiting time of the first customer in queue is

$$[\texttt{eqC1}] P(W_1 > t) = \Pi_{i=1}^{s}(1 - G_i(t|x_i)) \ , \tag{6.18}$$

which is easily calculated via

$$[\texttt{eqC2}] \log P(W_1 > t) = \sum_{i=1}^{s} \log(1 - G_i(t|x_i)) \ . \tag{6.19}$$

We can further approximate W_1 by an exponential distribution

$$[\texttt{eqC3}] P(W_1 > t) \approx e^{-\mu_1 t} \ , \quad t \geq 0 \ , \tag{6.20}$$

where μ_1 is obtained from (6.19) via

$$[\texttt{eqC4}] \mu_1 \approx \frac{\log P(W_1 > t_0)}{t_0} \tag{6.21}$$

for some appropriate t_0. This approximation is supported by extreme-value limits in the i.i.d. case, see Leadbetter, Lindgren and Rootzén [19] and Resnick [22]. As a supporting regularity condition, we assume that $G_i(t)$ and thus $G_i(t|x_i)$ has a positive density on the entire half line. We can then think of the initial departure process as a Poisson process with rate μ_1, so that W_k has approximately a gamma distribution with

$$[\texttt{eqC5}] EW_k = \frac{k}{\mu_1} \quad \text{and} \quad VarW_k = \frac{k}{\mu_1^2} \ . \tag{6.22}$$

Approximation (6.22) is also natural to use when we are only given the mean remaining service times of all s customers in service, say m_i, $1 \leq i \leq s$. Then we can let the rate μ_1 in (6.21) be

$$[\texttt{eqC6}] \mu_1 = \sum_{i=1}^{s} (1/m_i) \ . \tag{6.23}$$

If we assume that the service-time cdf's $G_i(t|x_i)$ are actually exponential with mean m_i, then the exact formula (6.19) reduces to (6.23).

The delay prediction analysis above is intended for the case of relatively large s and relatively small k, e.g., $s = 100$ and $k = 30$ (number in system $s+k = 130$). For the reverse situation (smaller s and larger k), different methods become more appropriate. Then the customers initially in service tend to play a smaller role. When $k \ll s$, it is natural to use a simple modification of (6.1) to take account of the non-identically distributed general cdf's G_i with means m_i and variances σ_i^2; i.e.,

$$[\texttt{eqW1}] EW \approx \frac{1}{s} \sum_{i=1}^{k+1} m_i \quad \text{and} \quad VarW \approx \frac{1}{s^2} \sum_{i=1}^{k+1} \sigma_i^2 \ , \tag{6.24}$$

where the s customers in queue are indexed first, followed by the first $s-k+1$ customers in service. We obtain (6.24) by approximately W by the sum of the first $k+1$ service times divided by s. In the i.i.d. case, the waiting time is the $(k+1)^{st}$ arrival time in the superposition of s i.i.d. renewal processes. Asymptotically as k increases, the mean and variance are as in (6.24); e.g., apply Theorem 6 of Glynn and Whitt [11].

Even if the elapsed service times are not available, we can do better than (6.1)–(6.3) for non-exponential distributions. Then, instead of (6.4), we would use the service-time stationary-excess cdf

$$\texttt{[eq613]} G_{ie}(t) = \frac{1}{m} \int_0^t [1 - G_i(u)] du \ , \quad t \geq 0 \ . \tag{6.25}$$

As noted in Duffield and Whitt [8], in the infinite-server model the residual service times, conditional on the number of busy servers, are distributed *exactly* as (6.25), so there is a theoretical basis for (6.25).

The importance of these alternatives to the exponential formulas in (6.1)–(6.3) clearly increases as the service-time distribution differs more from an exponential distribution. The difference is dramatic when the service-time distribution is a long-tail distribution such as the Pareto distribution. Indeed, suppose that $Y(a,b)$ has the Pareto cdf $G(t) = 1 - (1+bt)^{-a}\ , t \geq 0$. Let $Y_x(a,b)$ have the conditional cdf $G(t|x)$. Then, by Theorem 8 of [8], $Y_x(a,b)$ is distributed the same as $(1+bx)Y(a,b)$. Hence the mean residual residual life is approximately proportional to the age. Hence, in this setting the age can greatly help in predicting the residual life.

It is significant that the delay prediction method in (6.4)–(6.17) does not depend much on the BD model structure. For example, it can be used for non-Poisson and non-homogeneous arrival processes. The most critical assumption is that service times are independent of the remaining system state. However, experience shows that dependence can occur between service times and system state. Human servers may speed up or slow down under heavier loads. Even computer servers may behave in this way, e.g., service times in database systems tend to increase under higher loads. To investigate this phenomenon, when estimating the service-time cdf, the service-time data can be grouped according to the number in system when service starts. With no system-state influence, this extra variable should not alter the estimation. However, if it does, then the service-time cdf G_i in (6.4)–(6.12) can be made to depend on the state $s+k$ seen by the arrival whose delay we are trying to predict. It is often significant just to distinguish two cases: when all servers are busy and when they are not.

We have focused on the delay before beginning service, but interest may instead be focused on delay until completing service. Assuming that the service time of each customer in queue is independent of his waiting time to begin service, the distribution of the time to complete service is naturally estimated by the convolution of the two estimated component distributions. Similarly, the estimated mean is simply the sum of the component estimated means.

7. Estimating the Balking Parameters

[sec7] In this section we consider how to estimate the balking parameters α and β in Section 3 and how to validate the model. For background on standard procedures for estimating parameters in BD models, see Basawa and Prakasa Rao [3] and references cited there.

For $0 \le k \le r-1$, let $A_k(t)$ be the number of arrivals finding $s+k$ customers in the system upon arrival and let $J_k(t)$ be the number of these arrivals to join the queue in an operation of the system over a time interval $[0,t]$. (The number balking is thus $A_k(t) - J_k(t)$.) Under the model assumptions, as $t \to \infty$, the ratio will converge as the sampling period grows, i.e.,

$$\text{[eq701]}\, R_k(t) \equiv \frac{J_k(t)}{A_k(t)} \to \eta \equiv (1-\beta)\left(\frac{s\mu}{s\mu+\alpha}\right)^{k+1} \quad \text{as} \quad t \to \infty \ , \tag{7.1}$$

so that

$$\text{[eq702]} - \log R_{k-1}(t) \to -\log(1-\beta) - k \log\left(\frac{s\mu}{s\mu+\alpha}\right) \quad \text{as} \quad t \to \infty \ . \tag{7.2}$$

Moreover, under the model assumptions, conditional on $A_k(t)$, $J_k(t)$ has a binomial distribution with parameters $n = A_k(t)$ and $p = \eta$ in (7.1). Hence, we propose estimating the parameters α and β by performing a linear regression with the variables $-\log R_{k-1}(t)$, $k \ge 1$, i.e., we find the best linear fit

$$\text{[eq703]} - \log R_{k-1}(t) = \hat{a}_1 + \hat{a}_2 k \ . \tag{7.3}$$

We then estimate α and β by $\hat{\alpha}$ and $\hat{\beta}$, where

$$\text{[eq704]} - \log(1-\hat{\beta}) = \hat{a}_1 \quad \text{and} \quad -\log\left(\frac{s\mu}{s\mu+\alpha}\right) = \hat{a}_2 \ , \tag{7.4}$$

so that

$$\text{[eq705]}\, \hat{\beta} = 1 - e^{-\hat{a}_1} \tag{7.5}$$

and

$$\text{[eq706]}\, \hat{\alpha} = s\mu(e^{\hat{a}_2} - 1) \ . \tag{7.6}$$

By (7.2), these estimators of α and β are consistent (converge as $t \to \infty$). The degree to which a linear fit in (7.3) is appropriate also indicates the quality of the model fit.

When the fit is not good, we should question whether T has an exponential cdf. More generally, we could directly estimate the probability $q_k^* \equiv P(T > ES_k) \equiv 1 - H(ES_k)$ by

$$[\text{eq707}]\, \hat{q}_k^* = R_k(t) \ , \quad 0 \le k \le r-1 \ . \tag{7.7}$$

A disadvantage of (7.7) for prediction is that it yields r parameters instead of only 2. However, from (7.7) we obtain an estimate of the cdf H at r points, because $q_k^* = H((k+1)/s\mu)$, $\quad 0 \le k \le r-1$.

More generally, with data, it is natural to consider other two-parameter or three-parameter models for non-balking. For example, instead of $(1-\beta)\gamma^{k+1}$ in (7.1), we might consider $(1-\beta)(k+1)^{-\gamma}$. If we do estimate the balking probabilities in state $s+k$ for each k, then it is natural to impose a monotonicity condition, exploiting the condition that the balking probability should be increasing in k. See Barlow, Bartholomew, Bremner and Brunk [2] for appropriate statistical methods.

8. Estimating the Reneging Rate

[sec8] In this section we consider how to estimate the reneging rate α in Section 2 or δ (assuming $\delta_k = \delta$) in Section 3. As noted at the end of Section 5, the average conditional time to abandon $E(A|R)$ for the model in Section 2 is often substantially less than $1/\alpha$, the reciprocal of the reneging rate. As an estimator $\hat{\alpha}$ for α, we propose that value of α, with the other elements of the parameter tuple $(\lambda, \mu, \alpha, \beta, s, r)$ that yields the observed estimate for the mean $E(A|R)$; i.e., we directly estimate $E(A|R)$ by looking at the sample mean of the reneging times and then we apply the BD model to find that value of α that yields the estimate. The most important point is not to confuse $E(A|R)$ with $1/\alpha$.

Alternatively, we could estimate the long-run reneging rate by its sample mean, and then estimate α by the value that yields the observed sample average reneging rate. By Theorem 4.4, the long-run reneging rate is always increasing in α, so that the search is not difficult to perform, e.g., by bisection search. This estimation procedure can also be used when there is reneging even when delays are predicted.

When the service provider announces delay predictions to each arrival, it is possible that the reneging behavior depends on the initial state. To confirm the delay predictions and to understand the reneging behavior, it is good to monitor the outcomes starting with each initial state $s+k$ for $k \ge 0$. Reneging events well before the anticipated waiting time $(k+1)/s\mu$ represent an unwillingness to wait for the predicted time. Reneging events after the anticipated waiting time $(k+1)/s\mu$ represent a failure to accurately predict the delay and associated customer dissatisfaction.

9. Occasional Extra Long Service Times

[sec9] In this section we propose some simple methods to describe the impact of occasional extra long service times. The delay prediction methods in Section 6 should already cover this case adequately. Now we are primarily concerned with modifications to the model in Section 3 to produce appropriate approximate modified performance predictions. Our idea is to represent the special service times as server vacations or server interruptions. Since these service times are unusually long, they occur in a longer time scale. Thus, it is natural to represent these service times as special high-priority customers that occasionally require servers. Moreover, since the special service times are unusually long, it should be reasonable to treat the remaining customers by averaging the steady-state distributions associated with the various possible numbers of available servers.

Hence, we first model the long service times by an M/G/∞ model. The steady-state number of servers occupied with these special customers thus has a Poisson distribution with mean equal to $m_L \equiv \lambda_L/\mu_L$, where λ_L is the arrival rate and μ_L^{-1} is the mean of these special long service times. We are assuming that the total offered load of these special customers, m_L, is sufficiently small that the chance that all servers are busy serving only them is negligible. Because of the insensitivity of the M/G/∞ model, the service-time distribution beyond the mean plays no role at this point.

We can then consider the original model, where the number of servers is random (but fixed for all time) having the value $s - N_L$, where N_L has a Poisson distribution with mean m_L. That is, we consider the BD model in Section 3, where the number of servers is $s - k$. The arrival rate λ and mean service time μ^{-1} must be appropriately reduced to account for the removal of the especially long service times. For each $k \leq m + r\sqrt{m}$, say, we compute the steady-state distributions for the BD model with $s - k$ servers. The performance measures for the models with $s - k$ servers can then be averaged with regard to the Poisson probabilities of k servers being busy serving the long service times, but it may be more revealing to look at the conditional performance measures for fixed k, given those k whose likelihood is considered sufficiently large. Tables and plots of *both* the probability of k servers being used by the long-service-time customers and the conditional performance measures for the remaining customers given $s - k$ servers, as a function of k, should provide useful insight.

10. Coping With Other Model Deviations

[sec10] We conclude by briefly discussing other possible deviations from the basic BD model and how they might be coped with. Serious investigations of these procedures represent topics for future research.

Time Dependence. Perhaps the most common difficulty is that the arrival process can be nonstationary. In many applications a reasonable model for the arrival process is a nonhomogeneous Poisson process with deterministic arrival-rate function $\lambda(t)$ that varies over time; e.g., see Chapter 6 of Hall [14]. The service-time distribution may be time-dependent as well. One approach to this complication is to apply numerical methods to solve the time-dependent BD process, obtained by working with $\lambda(t)$ and $\mu(t)$ instead of λ and μ. A specific algorithm based on a discrete-time approximation is given in Davis, Massey and Whitt [6]. References are also cited there to sources applying the related Runge-Kutta methods to numerically solve the ordinary differential equations.

A simple approximation for the time-dependent distribution of the time-dependent BD process is the pointwise stationary approximation (PSA), which is the steady-state distribution of the BD process calculated in terms of the arrival-rate and service-rate functions $\lambda(t)$ and $\mu(t)$ as a function of time t. If $\lambda(t)$ varies significantly over time, then the PSA is often a far better description than the BD model with the long-run average arrival and service rates; e.g., see Green and Kolesar [12]. The PSA is also asymptotically correct as the arrival and service rates increase which corresponds to the rates changing more slowly; see Whitt [29]. In other words, the steady-state analysis here is directly applicable as a reasonable approximation when the arrival and service rates fluctuate if it is applied over suitable subintervals over which these functions do not change much. The estimated rates are then averages over these subintervals.

A complication where time-dependence is recognized is that it becomes necessary to estimate the functions $\lambda(t)$ and $\mu(t)$ instead of the single parameters λ and μ. Appropriate data smoothing is thus often required.

Non-Exponential Service-Time Distributions. We have considered how to exploit non-exponential service-time distributions to predict delays in Section 6. However, non-exponential service-time distributions also will tend to invalidate the BD model predictions. The congestion is likely to be greater (less) if the service-time distribution is more (less) variable than exponential. The impact of a non-exponential service-time distribution can be at least roughly estimated by examining its impact on the related M/G/s/∞ pure-delay model; e.g., see Whitt [30] and references cited there for simple approximations.

The impact of a non-exponential service-time distribution should be negligible if the arrival process is Poisson and the probability that all servers are busy is small, because the M/G/s/0 and M/G/∞ models have the insensitivity property. However, the steady-state behavior conditional on all servers being busy should be significantly affected by the service-time distribution beyond its mean.

In Section 8 we proposed a way to study the impact of a few exceptionally long service times. If the service-time distribution can be regarded as approximately exponential after removing such exceptionally long service times, then the modified BD analysis in Section 8 should be successful.

Similarly, if there is an excess of customers with very short service times, then they could be ignored. The resulting lower arrival rate and higher mean service time of the remaining customers may yield more accurate descriptions, assuming a BD model based on the approximate exponential distribution.

Non-Poisson Arrival Processes. In many settings, the Poisson arrival process (possibly non-homogeneous) is natural, representing the result of many different customers making independent decisions. However, if the Poisson property is not nearly realistic, then the BD predictions can be far off. Non-Poisson processes arise naturally when the arrival process is itself an overflow process from another group of servers.

One way to approximately cope with non-Poisson stationary arrival processes is to substitute time-dependence or state-dependence for the stochastic dependence in the actual arrival process. The use of time-dependence is to reverse the approximation procedure discussed in Massey and Whitt [20]. In our setting with balking and reneging, the time-dependent birth-and-death process may be substantially easier to analyze than the stationary model with a non-Poisson arrival process.

Alternatively, we can try to approximately represent stochastic variability by a state-dependent arrival rate. In particular, we could use the Bernoulli-Poisson-Pascal (BPP) model in which the arrival rate λ is replaced by the linear function $\lambda_k = \alpha + \beta k$ for $k \geq 0$; see Delbrouck [7] and Choudhury, Leung and Whitt [5]. The less bursty binomial case corresponds to $\beta < 0$, while the more busty Pascal case corresponds to $\beta > 0$.

All these analytical approximations can be substantiated by computer simulation.

Acknowledgment. I thank Avishai Mandelbaum of the Technion for helpful pointers to the literature.

References

[1] J. Abate and W. Whitt, Numerical Inversion of Laplace Transforms of Probability Distributions, *ORSA J. Computing* 7 (1995) 36–43.

[2] R. E. Barlow, D. J. Bartholomew, J. M. Bremner and H. D. Brunk, *Statistical Inference Under Order Restrictions*, Wiley, New York, 1972.

[3] I. V. Basawa and B. L. S. Prakasa Rao, *Statistical Inference for Stochastic Processes*, Academic Press, New York, 1980.

[4] O. J. Boxma and P. R. de Waal, Multiserver Queues with Impatient Customers, *Proceedings ITC 14*, J. Labetoulle and J. W. Roberts (eds.), North-Holland, Amsterdam, 1995, 743–756.

[5] G. L. Choudhury, K. K. Leung and W. Whitt, An Inversion Algorithm to Compute Blocking Probabilities in Loss Networks with State-Dependent Rates, *IEEE/ACM Trans. Networking* 3 (1995) 585–601.

[6] J. L. Davis, W. A. Massey and W. Whitt, Sensitivity of the Service-Time Distribution in the Nonstationary Erlang Loss Model, *Management Sci.* 41 (1995) 1107–1116.

[7] L. E. N. Delbrouck, A Unified Approximate Evaluation of Congestion Functions for Smooth and Peaky Traffic, *IEEE Trans. Commun.* COM29 (1981) 85–91.

[8] N. G. Duffield and W. Whitt, Control and Recovery from Rare Congestion Events in a Large Multi-Server System, *Queueing Systems* (1997) to appear.

[9] G. Falin, A Survey of Retrial Queues, *Queueing Systems* 7 (1990) 127–167.

[10] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. II, second edition, Wiley, New York, 1971.

[11] P. W. Glynn and W. Whitt, Ordinary CLT and WLLN Versions of $L = \lambda W$. *Math. Oper. Res.* 13 (1988) 674–692.

[12] L. Green and P. Kolesar, The Pointwise Stationary Approximating for Queues with Nonstationary Arrivals, *Management Sci.* 37 (1991) 84–97.

[13] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, second edition, Wiley, New York, 1985.

[14] R. W. Hall, *Queueing methods for Services and Manufacturing*, Prentice Hall, Englewood Cliffs, NJ, 1991.

[15] D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research*, Vol. I, McGraw-Hill, New York, 1982.

[16] M. K. Hui and D. K. Tse, What to Tell Customers in Waits of Different Lengths: An Integrative Model of Service Evaluation. *J. Marketing* 60 (1996) 81–90.

[17] K. L. Katz, B. M. Larson and R. C. Larson, Prescription for the Waiting-in-Line Blues: Entertain, Enlighten and Engage. *Sloane Management Review* 32 (1991) 44–53.

[18] F. P. Kelly, Loss Networks, *Ann. Appl. Prob.* 1 (1991) 319–378.

[19] M. R. Leadbetter, G. Lindgren and H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York, 1983.

[20] W. A. Massey and W. Whitt, Stationary-Process Approximations for the Nonstationary Erlang Loss Model, *Opns. Res.* 44 (1996) 976–983.

[21] D. M. Rappaport, Key Role of Integration in Call Centers, *Business Communications Review*, July 1996, 44–48.

[22] S. I. Resnick, *Extreme Values, Regular Variation and Point Processes*, Springer-Verlag, New York, 1987.

[23] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunication Networks*, Springer, New York, 1995.

[24] M. Shaked and J. G. Shanthikumar, *Stochastic Orders and Their Applications*, Academic Press, New York, 1994.

[25] D. R. Smith and W. Whitt, Resource Sharing for Efficiency in Traffic Systems. *Bell System Tech. J.* **60** (1981) 39–55.

[26] S. Taylor, Waiting for Service: The Relationship Between Delays and Evaluations of Service. *J. Marketing* 58 (1994) 56–69.

[27] W. Whitt, Comparing Counting Processes and Queues, *Adv. Appl. Prob.* 13 (1981) 207–220.

[28] W. Whitt, Blocking When Service Is Required from Several Facilities Simultaneously, *AT&T Tech. J.* 64 (1985) 1807–1856.

[29] W. Whitt, The Pointwise Stationary Approximation for $M_t/M_t/s$ Queues Is Asymptotically Correct As the Rates Increase, *Management Sci.* 37 (1991) 307–314.

[30] W. Whitt, Approximations for the GI/G/m queue, *Production and Operations Management* 2 (1993) 114–161.

[31] R. W. Wolff, *Stochastic Modelling and the Theory of Queues*, Prentice Hall, Englewood Cliffs, NJ, 1989.