Double-Sided Batch Queues with Abandonment: Modeling Crossing Networks

Philipp Afèche, Adam Diamant, Joseph Milner
University of Toronto, Rotman School of Management, 105 St. George Street, Toronto, Ontario, M5S 3E6
afeche@rotman.utoronto.ca, adam.diamant09.rotman.utoronto.ca, jmilner@rotman.utoronto.ca

We model a double-sided queue with batch arrivals and abandonment. We consider two types of customers, those with some patience and those that depart immediately if their order is not filled. The model is particularly applicable to a class of alternative trading systems called crossing networks that are increasingly important in the operation of modern financial markets. We derive stability conditions and the steady-state distribution for the queueing model assuming that units brought by customers abandon the queue at a constant rate. We present results on the expected system time and fill rate for this queueing system. Typically, queueing models assume abandonment of units after some customer specific deadline. We derive the expected results for a customer with a given deadline who departs the modeled queue and compare our results to those in a simulation.

Key words: double-sided queues, balking and reneging, batch arrivals, level crossing, crossing networks

1. Introduction

We consider a model of a double-sided queue with batch arrivals and abandonment. Two types of customers, patient and impatient, arrive to each side of the queue where only patient customers are willing to queue. Each customer brings a random number of units. The arrival process for each customer type is modeled as a side-specific, independent compound Poisson process where the jump size distribution (corresponding to order batch size) is exponentially distributed. Units in queue on one side of the queue are "served" by units from customers from the opposite side. Units from impatient customers depart immediately if not served. Any extra arriving units from patient customers queue, however, we assume that some of these queued units may abandon the system at a future time. We assume units abandon the system at a constant, side-dependent rate. This is in contrast to a process that tracks the tenure of individual customers in the system, as is common in some models of abandonment and have these customers depart if not served by some deadline. Doing so allows us to create a tractable model for which we can provide a closed-form characterization of the steady-state system behavior. We then derive expressions for system-level and customer-level performance measures such as mean system time and percentage of units served.

We compare the results from our model to simulations where customers (not units) abandon the system after some random deadline to validate the applicability of the model.

Our model is the first to present closed-form results for a double-sided queueing model with batch arrivals and abandonment. Previous work on such double-sided queues have considered batch queues (e.g., Kashyap (1966)) and unit arrivals with abandonment (e.g., Zenios (1999), Boxma et al. (2011)). Only Kim et al. (2010) have considered both batch queues and abandonment, however they do so only through a numerical simulation model. Further, we consider heterogenous customers with type- and side-dependent arrival rates and batch sizes, previously not considered.

We focus on this type of queue as it closely models the operations of a crossing network, one of several variants of increasingly important financial trading markets. Crossing networks and other "dark exchanges" (or dark pools) act as alternatives to the more familiar trading markets such as the New York Stock Exchange (NYSE) and NASDAQ. In dark exchanges, traders (customers) can submit buy and sell market orders that are hidden from other market participants. Information such as the stock being traded, whether the trader wishes to buy or sell a security, and the number of units being transacted, are not disclosed until after a trade is carried out. Successful trades (at least in the U.S.) are recorded to the national consolidated tape as over-the-counter transactions without detailed information, such as the parties involved, the exchange responsible for facilitating the trade, and the exact time of the transaction. The lack of publicly available information means that at no point can parties directly observe the volume or liquidity available in this trading venue, which is in direct contrast to transparent or "light" exchanges like the NYSE. In 2010, 13.27% of U.S. equities trading volume was transacted in dark exchanges which represented a 30.7% increase over 2009 (Schack and Gawronski 2011).

The most common dark exchange, crossing networks, are ones where traders, referred to below as customers, submit anonymous buy or sell orders for a particular security, along with the order size and the maximum transaction time which we refer to as the deadline. Customers do not see the state of the system. Orders from counter-parties are matched without any intervention from dealers, brokers or market makers in FIFO order. If sufficient liquidity is present in the crossing network, trades are carried out at a price exogenously given - for example the midpoint of the bid/ask spread derived from the National Best Bid and Offer (NBBO). If an order is not completely filled within the required time, the remainder of the order is canceled (the remaining shares may be submitted to a different exchange).

In many crossing networks, both market orders (an order to buy or sell stocks at the prevailing market price) and limit orders (an order to buy or sell a stock at a specific price or better) are accepted, which means that at any point in time, there may be an excess of orders that wish to buy or sell a stock. We only consider market orders as (1) the volume of limit orders at this time

is small compared with market orders in the crossing networks we are familiar with (ITG 2011), and (2) allowing limit orders significantly changes the behavior of the queue under consideration, and so is left for future work.

The remainder of the paper proceeds as follows. In Section 2, we review related research on double-sided queueing as well as the most recent developments in the modeling of trading markets. In Section 3, we introduce the queueing model. In Section 4 we derive the steady-state distribution for the model and its stability conditions, and provide formulae for the system's expected performance. In Section 5 we derive metrics on the performance of the system for customers with a known (possibly infinite) deadline. This is done to understand the experience of individual customers with specific attributes as opposed to system averages. We compare the results of our model to those of a simulation where customers arrive with deadlines in order to validate our assumption of an aggregate abandonment rate in Section 6. We discuss our results in Section 7 and present a number of directions for future research.

2. Literature Review

In this section we review research on double-sided queueing models, dam and insurance models which are related to our aggregate abandonment model, associated models on perishable inventory, and recent research on financial trading markets. Kendall (1951) introduced the double-sided queueing model using the example of taxis and customers independently arriving to a queueing point. Solution methods for the double-sided queueing model were introduced in the early 1960's by Dobbie (1961). Additional research in the area includes time-dependent arrival rates (Giveen 1963), bulk service (Kashyap 1966), and limited waiting space for both taxis and customers (Gaur and Kashyap 1973). More recently, the model has been applied to assembly facilities producing multi-part components. Parts arrive independently according to a renewal process, but the component can only be assembled when all the parts are present. A number of papers have discussed these assembly-like queues including Som et al. (1994) and Takahashi et al. (2000). The double-sided queueing model also has a number of applications in other areas including parallel processing, database concurrency control, communication protocols and inventory management.

Zenios (1999) and Boxma et al. (2011) consider double-sided queues with abandonment and apply the model to the organ transplant waiting list in the United States. Both organs and patients arrive independently to a queue, but may abandon as organs can expire and patients can die. Conolly et al. (2002) extends the basic model for a double-sided queue with impatience to include time-dependent behavior and state-dependent abandonment processes. A number of related models were considered, including ones where abandonment occurs on either side. These models focus on the arrival and departure of single-units only and do not take into account bulk-arrivals. Kim

et al. (2010) present a simulation model of a bulk-arrival, batch-service, double-sided queue with abandonment. However, they only provide a numerical procedure to find the state probabilities and various performance measures.

Our research is related to dam and insurance models in which the state of the system changes either by a jump or a continuous process (c.f. Asmussen (2003)). Recent papers by Perry et al. (1999), Perry et al. (2002) and Lee (2007) analyze dam processes whose system state can increase and decrease through jumps while also decreasing continuously. Similarly the insurance ruin literature has incorporated upward jumps (e.g., large increases in capital that are unforeseen) into the traditional model of downward jumps (e.g., insurance claims) and a continuous premium collected at a constant rate per unit time; see, e.g., Geng (2008). These papers focus on the transient properties of the dam or insurance process such as the time until the system is empty and the total amount of claims paid prior to ruin. Further, such models are naturally defined only on the half-line in comparison to a double-sided queue. In comparison, we calculate metrics such as the number of units in the system, the fill rate, and the average amount of time a unit spends in the system. Further, our system operates perpetually and does not condition on a termination event (i.e., ruin or overflow) which is the foundation of insurance research and dam models.

Our research is also related to the application of queueing theory to continuous review, perishable inventory systems as introduced by Graves (1982) and Kaspi and Perry (1983). In such work impatient customers are either served from inventory or queue, while inventory units are supplied by a continuous production process and expire after some fixed amount of time. This model has been modified to allow for multiple unit demand and random customer abandonment times (Kaspi and Perry 1984), items with random expirations (Perry 1985), and state-dependent arrival and departure rates with finite queue lengths (Perry and Stadje (1999) and Nahmias et al. (2004)). Goh et al. (1993) consider the case of geometrically distributed batch sizes. Recently, Guo et al. (2011) study a system where with a fixed, continuous replenishment policy on one side of the queue and customer demand on the other side arriving according to a pricing controlled compound Poisson process. In contrast, we consider stochastic arrivals on both sides of the queue. The latter is more in keeping with the motivating example of arrivals to a crossing network.

The literature on crossing networks and other dark pools is quite sparse. Hendershott and Mendelson (2002) and Ray (2010) study the conditions under which investors should use a dark pool versus a traditional trading venue. Ready (2009) studies the constituency of the traders in a dark exchange and its relation to market liquidity. Buti et al. (2011) study whether trading in dark pools has a detrimental effect on the market due to the decrease in liquidity and price improvement. Several papers from the industry, e.g., Sofianos (2007) and Mittal (2008), describe

the various flavors of dark pools and their effect on the trading system as a whole. No work has been undertaken to understand the *operation* of these trading systems.

In contrast, there has been considerable recent literature on the operation of a limit order book in a visible exchange, see, e.g., Parlour (1998), Foucault et al. (2005), Rosu (2009), Maglaras and Moallemi (2011). They study the strategic behavior of traders given the amount of liquidity available at each price. While limit orders are accepted at some crossing networks, the order book is *not* visible. Thus these models are not directly applicable to crossing networks as visible exchanges have different dynamics compared with their visible counterparts (c.f. Buti et al. (2011), Zhu (2012)).

3. The Model

In this section, we introduce the double-sided queueing model with batch arrivals and abandonment. We designate the two sides of the queue as a and b. Briefly, we model a system with Poisson arrivals bringing an exponentially distributed number of units to both sides. Queued units are served FCFS by arrivals from the opposite side. Units that do not serve queued units, themselves then queue. Let $Y(t)$ be the stochastic process associated with the queue length of the double-sided queue and, if it exists, let Y be the random variable of the stationary queue length with CDF $F(y)$. That is, let $\lim_{t\to\infty} P(Y(t) \leq y) = F(y)$ if $Y(t)$ converges in distribution. Let $f(y)$ be the associated probability density function. To simplify the description of the system behavior, let $y(t)$ be a realization of $Y(t)$, the number of units in the queue, where $y(t) < 0$ implies a queue of side-a units and $y(t) > 0$ implies a queue of side-b units.

The arrival process to the queue is as follows. We refer to 'side-a' or 'side-b' arrivals as appropriate. For each side, we consider two types of arrivals, patient and impatient; we denote the parameters pertaining to the impatient customers with a superscript I. Demand for each type and side arrives according to mutually independent compound Poisson processes. Let $N_a(t)$ be the Poisson process for the customers arrivals of side-a patient customers. Similarly define $N_b(t)$, $N_a^I(t)$, and $N_b^I(t)$. Let the associated arrival rates be λ_j, $j \in \{a, b\}$ for patient and λ_j^I, $j \in \{a, b\}$ for impatient customers. We assume that each patient (impatient) arrival brings an exponentially distributed number of units. Let X_{ai} be the number of units brought by the i^{th} side-a patient arrival. Similarly define X_{bi}, X_{ai}^I, and X_{bi}^I. Let the mean number of units brought be ω_j (ω_j^I), $j \in \{a, b\}$ for patient (impatient) customers. For notational convenience, let $\mu_j = 1/\omega_j$ and $\mu_j^I = 1/\omega_j^I$, $j \in \{a, b\}$.

The double-sided queue operates in FCFS fashion with units from an arriving customer on one side of the queue (say, side-a) serving any queued units on the other side (side-b). Side-a units not serving queued side-b units join the side-a queue if the arrival is a patient customer (and similarly for arriving side-b units). Units from impatient customers that do not serve queued units depart

immediately. We assume a one-for-one matching policy where an equal number of side-a units and side-b units exit the system, leaving the remainder to queue (again, if patient). When there is a queue of side-a units, the absolute queue length decreases at some rate $r_a(y) \geq 0$ (and similarly for side-b, $r_b(y) \geq 0$), representing units abandoning the system. Let

$$r(y) = -r_a(y)1_{y<0} + r_b(y)1_{y>0}.$$

where 1_A is the indicator function of event A. In this regard, the patient customers are not infinitely patient; their units abandon the queue after some time. Throughout the paper we assume $r_i(y) = k_i$ for some constants $k_i \geq 0$, $i \in \{a, b\}$.

For counting process $N(t)$, let $dN(t)/dt$ be the Dirac delta function for process $N(t)$. Then the process $Y(t)$, given some initial position y_0, is defined by the differential equation

$$\begin{aligned}\frac{dY(t)}{dt} = &- X_{aN_a(t)}\frac{dN_a(t)}{dt} + X_{bN_b(t)}\frac{dN_b(t)}{dt} \\ &- \min\left[X^I_{aN^I_a(t)}, (Y(t))^+\right]\frac{dN^I_a(t)}{dt} + \min\left[X^I_{bN^I_b(t)}, (-Y(t))^+\right]\frac{dN^I_b(t)}{dt} - r(Y(t)).\end{aligned}$$

The first two terms give the jump process for the side-a and $-b$ patient arrivals, the third and fourth terms, the jumps for the impatient arrivals, and the last term expresses the abandonment rate. A sample path depicting the $Y(t)$ process is shown in Figure 1. If the system converges, there is a non-zero probability that the system will be empty; let $\mathbb{P}_0$ be this probability.

3.1. Model Justification with Respect to Crossing Networks

The relationship between the model and the behavior of a crossing network rests on assumptions regarding the arrival process, the order size distribution and the abandonment process. Previous work, for example, Foucault et al. (2005) and Rosu (2009), model the arrival process in financial markets as Poisson, noting the acceptability of doing so under stable market conditions. Similarly, Gopikrishnan et al. (2000) and Maslov and Mills (2001) demonstrate that the order size distribution in transparent markets is well described by either a power law, or an exponential form. However, no work has been done to analyze the order size distribution for crossing networks, primarily as a result of their recent advent and the anonymity of the participants.

We consider two abandonment processes. For patient customers, we assume that they depart the system after queueing for some time. This class of customers provide liquidity in crossing networks and are necessary for their functioning. Typically such customers are those that have made a commitment to hold or divest from a security. As such they will hold their orders in the market for some period of time before withdrawing them. As discussed above, we model the aggregate abandonment of units from the queue rather than each individual's abandonment to maintain

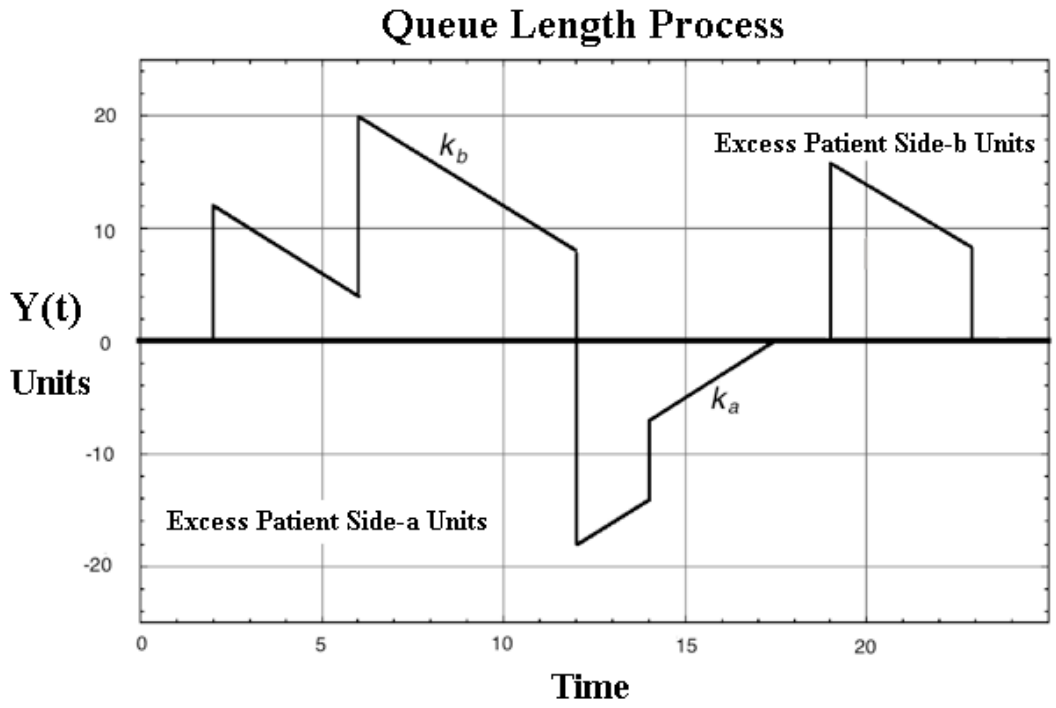

Figure 1 A typical sample path of the queue length process. The positive half-line represent excess side-b units while the negative half-line represent excess side-a units where $r_i(y) = k_i$, $i = \in \{a, b\}$.

tractability. Doing so results in system behavior very similar to simulation results for the case where customers withdraw all unserved units at the same time.

Impatient customers model a class of customers often seen within crossing networks who submit what is known as immediate-or-cancel orders (IOC). Such customers test the market by submitting orders (often of a relatively small size) to determine if a queue exists on one side or the other, withdrawing their orders almost immediately if not filled. By doing so, they may gain information from the crossing network that by its nature is obscured. This information may be exploited by taking simultaneous positions in both light and dark markets.

4. Stationary Distribution

In this section we derive the steady-state distribution for the $Y(t)$ process using techniques from level-crossing theory. We then provide expectations for the mean fill rate and system time.

4.1. Derivation

PROPOSITION 1. *$f(y)$ is given by the solution to the following differential equations:*

For $y > 0$,

$$f'''(y)r(y) + f''(y)\Big(3r'(y) + \big(\mu_b - \mu_a - \mu_a^I\big)\, r(y) - \big(\lambda_a + \lambda_b + \lambda_a^I\big)\Big) \tag{1}$$

$$
\begin{aligned}
&+ f'(y)\Big(3r''(y) + 2\left(\mu_b - \mu_a - \mu_a^I\right) r'(y) + \left(\mu_a^I\mu_a - \mu_b\mu_a^I - \mu_b\mu_a\right) r(y)\Big) \\
&+ f'(y)\lambda_a^I\left(\mu_a - \mu_b\right) + \lambda_b\left(\mu_a + \mu_a^I\right) + \lambda_a\left(\mu_a^I - \mu_b\right)\Big) \\
&+ f(y)\Big(r'''(y) + \left(\mu_b - \mu_a - \mu_a^I\right) r''(y) + \left(\mu_a\mu_a^I - \mu_b\mu_a^I - \mu_b\mu_a\right) r'(y) + \mu_a\mu_a^I\mu_b r(y)\Big) \\
&+ f(y)\Big(\lambda_a\mu_b\mu_a^I + \lambda_a^I\mu_b\mu_a - \lambda_b\mu_a\mu_a^I\Big) \\
&= 0,
\end{aligned}
$$

and for $y < 0$,

$$
\begin{aligned}
&f'''(y)r(y) + f''(y)\Big(3r'(y) + \left(\mu_b^I + \mu_b - \mu_a\right) r(y) - \left(\lambda_a + \lambda_b + \lambda_b^I\right)\Big) \\
&+ f'(y)\Big(3r''(y) + 2\left(\mu_b^I + \mu_b - \mu_a\right) r'(y) + \left(\mu_b^I\mu_b - \mu_a\mu_b^I - \mu_a\mu_b\right) r(y)\Big) \\
&+ f'(y)\Big(\lambda_b^I\left(\mu_a - \mu_b\right) - \lambda_b(\mu_b^I - \mu_a) - \lambda_a(\mu_b^I + \mu_b)\Big) \\
&+ f(y)\Big(r'''(y) + \left(\mu_b^I + \mu_b - \mu_a\right) r''(y) + \left(\mu_b^I\mu_b - \mu_a\mu_b^I - \mu_a\mu_b\right) r'(y) - \mu_a\mu_b\mu_b^I r(y)\Big) \\
&+ f(y)\Big(\lambda_b\mu_a\mu_b^I + \lambda_b^I\mu_a\mu_b - \lambda_a\mu_b\mu_b^I\Big) \\
&= 0.
\end{aligned}
\tag{2}
$$

(All proofs appear in the Appendix.)

Differential equations (1) and (2) are third-order and homogeneous, with constant coefficients over the half-line, and each has a boundary at zero. Three boundary conditions determine their solution.

The first is the normalization condition:

$$\int_{-\infty}^{\infty} dF(y) = 1. \tag{3}$$

The second and third are flow balance conditions. Let $s_a(y)$ denote the expected number of units served by a patient side-a customer given $y > 0$ units are in the queue upon arrival. Then,

$$s_a(y) = \int_0^y \mu_a z e^{-\mu_a z}\, dz + \int_y^{\infty} \mu_a y e^{-\mu_a z}\, dz = \frac{1}{\mu_a}(1 - e^{-\mu_a y}).$$

Let $s_a^I(y)$ denote the expected number of units served by an impatient side-a customer given $y > 0$ units are in the queue upon arrival. Then,

$$s_a^I(y) = \frac{1}{\mu_a^I}(1 - e^{-\mu_a^I y}).$$

Similarly,

$$s_b(y) = \frac{1}{\mu_b}(1 - e^{\mu_b y}) \;\text{ and }\; s_b^I(y) = \frac{1}{\mu_b^I}(1 - e^{\mu_b^I y}),$$

are respectively, the number of units served by a patient and impatient side-b customer given $y<0$ units are in queue. The expected numbers of units served are

$$\bar{s}_a \triangleq \int_0^\infty s_a(y)f(y)\,dy \quad \text{and} \quad \bar{s}_b \triangleq \int_{-\infty}^0 s_b(y)f(y)\,dy,$$

for patient side-a and side-b customers, respectively. For impatient side-a and side-b customers the expected values are

$$\bar{s}_a^I \triangleq \int_0^\infty s_a^I(y)f(y)\,dy \quad \text{and} \quad \bar{s}_b^I \triangleq \int_{-\infty}^0 s_b^I(y)f(y)\,dy.$$

In steady-state the inflow rate of patient units into the system equals their outflow rate. The inflow rate is given by $\lambda_a\omega_a$ for patient side-a units and $\lambda_b\omega_b$ for patient side-b units. The outflow rate of the system is given by the summed service rate from both customer types, plus the rate units exit by abandonment. Recall that an equal number of units of patient side-a and side-b are served by each arrival. Thus, we have the following two flow balance equations:

$$\lambda_a\omega_a = \lambda_a\bar{s}_a + \lambda_b\bar{s}_b + \lambda_b^I\bar{s}_b^I + \int_{-\infty}^0 r(y)f(y)\,dy, \tag{4}$$

$$\lambda_b\omega_b = \lambda_a\bar{s}_a + \lambda_b\bar{s}_b + \lambda_a^I\bar{s}_a^I + \int_0^\infty r(y)f(y)\,dy. \tag{5}$$

Theorem 1 gives the closed-form solution for the steady-state density $f(y)$ for the case of $r(y) = k_b 1_{\{y>0\}} - k_a 1_{\{y<0\}}$.

THEOREM 1. *Under boundary conditions (3), (4) and (5), and $r(y) = k_b 1_{\{y>0\}} - k_a 1_{\{y<0\}}$, the double-sided queue is stable if and only if*

$$-k_a - \lambda_b^I\omega_b^I < \lambda_b\omega_b - \lambda_a\omega_a < \lambda_a^I\omega_a^I + k_b. \tag{6}$$

The stationary distribution of the queue length is then:

$$f(y) = \begin{cases} U_a\, e^{y\Theta_a} & y<0, \\ \mathbb{P}_0 & y=0, \\ U_b\, e^{-y\Theta_b} & y>0, \end{cases} \tag{7}$$

where

$$\begin{aligned}\delta = {} & \mu_a\mu_b\left(\Theta_a\left(\Theta_b\left(\lambda_b^I\lambda_a + \lambda_a^I\left(\lambda_b+\lambda_b^I\right)\right) + \lambda_a^I\left(\lambda_b+\lambda_b^I\right)\mu_a + \lambda_b^I\lambda_a\mu_a^I\right) + \lambda_a^I\lambda_b^I\mu_a\mu_b + \lambda_b^I\lambda_a\mu_a^I\mu_b\right. \\ & +\lambda_a^I\lambda_b\mu_b^I\mu_a + k_a\left(\Theta_b\left(\lambda_a+\lambda_a^I\right) + \lambda_a^I\mu_a + \lambda_a\mu_a^I + k_b\left(\Theta_b+\mu_a\right)\left(\Theta_b+\mu_a^I\right)\right)\left(\Theta_a+\mu_b\right)\left(\Theta_a+\mu_b^I\right) \\ & \left.+k_b\left(\Theta_b+\mu_a\right)\left(\Theta_b+\mu_a^I\right)\left(\Theta_a\left(\lambda_b+\lambda_b^I\right) + \lambda_b^I\mu_b + \lambda_b\mu_b^I\right) + \Theta_b\left(\left(\lambda_a+\lambda_a^I\right)\lambda_b^I\mu_b + \lambda_a^I\lambda_b\mu_b^I\right)\right),\end{aligned}$$

$$U_a = \frac{\Theta_a\lambda_a\left(\Theta_a+\mu_b\right)\left(\Theta_a+\mu_b^I\right)\left(\lambda_a^I\mu_a\mu_b + \Theta_b\left(\left(\lambda_a+\lambda_a^I\right)\mu_b - \lambda_b\mu_a\right) + \left(\lambda_a\mu_b - \lambda_b\mu_a\right)\mu_a^I + k_b\mu_b\left(\Theta_b+\mu_a\right)\left(\Theta_b+\mu_a^I\right)\right)}{\delta},$$

$$U_b = \frac{\Theta_b \lambda_b \left(\Theta_b + \mu_a\right)\left(\Theta_b + \mu_a^I\right)\left(\lambda_b^I \mu_a \mu_b + \Theta_a\left(\left(\lambda_b + \lambda_b^I\right)\mu_a - \lambda_a \mu_b\right) + \left(\lambda_b \mu_a - \lambda_a \mu_b\right)\mu_b^I + k_a \mu_a \left(\Theta_a + \mu_b\right)\left(\Theta_a + \mu_b^I\right)\right)}{\delta},$$

$$\mathbb{P}_0 = 1 - \frac{U_a}{\Theta_a} - \frac{U_b}{\Theta_b},$$

and Θ_a and Θ_b are non-negative constants given the proof that are explicit functions of the system parameters.

Theorem 1 states that the steady-state distribution of the queue length decreases away from $y = 0$ at an exponential rate. Further, there is an atom of probability ($\mathbb{P}_0$) that the system will empty as long as it is stable. Stability is guaranteed if and only if $-k_a - \lambda_b^I \omega_b^I < \lambda_b \omega_b - \lambda_a \omega_a < \lambda_a^I \omega_a^I + k_b$. Define the net-inflow to the queue as $\lambda_b \omega_b - \lambda_a \omega_a$ and the potential excess outflow on side-i as $\zeta_i = k_i + \lambda_j^I \omega_j^I$, for $i \in \{a, b\}$, $j \neq i$. So ζ_i is the rate of abandonment on side-i of the queue plus the rate that impatient units arrive on the other side (side-j) of the queue. Then the system is stable if the net-inflow is exceeded by the potential excess outflow for both sides of the queue.

To further understand the system's behavior, we introduce two alternate systems and compare their behavior to the original system. In the first alternate system, there is no abandonment, but the arrival rate of impatient units on side-i equals ζ_j. Let this "Impatient-Only" system be designated by a double-hat. That is, $\hat{\hat{k}}_i = 0$, $\hat{\hat{\lambda}}_i^I \hat{\hat{\omega}}_i^I = \zeta_j = k_j + \lambda_i^I \omega_i^I$ for $i = \{a, b\}$, $j \neq i$ where $\hat{\hat{\omega}}_i^I = \omega_i^I$ for $i \in \{a, b\}$. In the second system, there are no impatient customers, but the abandonment rate on side-i equals ζ_i. Let this "Abandonment-Only" system be designated by a double-bar. That is, $\bar{\bar{\lambda}}_i^I = \bar{\bar{\omega}}_i^I = 0$ and $\bar{\bar{k}}_i = \zeta_i = k_i + \lambda_j^I \omega_j^I$ for $i = \{a, b\}$, $j \neq i$. Let $\hat{\hat{\mathbb{P}}}_0$ be the probability of arriving to an empty system in the Impatient-Only system, and $\bar{\bar{\mathbb{P}}}_0$ be that for the Abandonment-Only system. We make two observations: (1) By appropriate choice of parameters, the steady-state distribution of a system with abandonment and/or impatient customers can be shown equivalent to that of a modified Abandonment-Only system; and (2) using the original system parameter values, we observe that P_0 is bracketed by $\hat{\hat{P}}_0$ and $\bar{\bar{P}}_0$.

To this end, in the proof of Theorem 1, we observe that for $y > 0$,

$$\begin{aligned} & r(y) f(y) + \lambda_a^I \int_y^\infty \exp\{-\mu_a^I (z - y)\}\, dF(z) \qquad (8) \\ & = \lambda_b \int_{-\infty}^y \exp\{-\mu_b (z - y)\}\, dF(z) - \lambda_a \int_y^\infty \exp\{-\mu_a (z - y)\}\, dF(z) \triangleq \text{net } b - \text{inflow across y} \end{aligned}$$

The right-hand side expresses the net in-flow of side-b units into the system above level y, while the left hand side provides the out-flow of the units through abandonment or through the arrival of impatient side-a customers. Stability is guaranteed when there is sufficient out-flow to balance the net in-flow. Let

$$\rho_b(y) \triangleq \frac{\lambda_a^I \int_y^\infty \exp\{-\mu_a^I (z - y)\}\, dF(z)}{f(y)}.$$

$\rho_b(y)$ expresses the rate of impatient units arriving and not serving customers at y. Because $f(y) = U_b \exp(-\Theta_b y)$ for $y > 0$,

$$\begin{aligned}\rho_b(y) &= \frac{\lambda_a^I \int_y^\infty \exp\{-\mu_a^I(z-y)\}\, dF(z)}{f(y)} \\ &= \frac{\lambda_a^I}{\mu_a^I + \Theta_b}.\end{aligned}$$

Observing $\rho_b(y)$ is independent of y (given $y > 0$) we can define $\kappa_a \triangleq \frac{\lambda_b^I}{\mu_b^I + \Theta_a}$. Similarly, defining $\kappa_b \triangleq \frac{\lambda_a^I}{\mu_a^I + \Theta_b}$, and letting $\rho(y) = -\kappa_a \mathrm{I}_{\{y<0\}} + \kappa_b \mathrm{I}_{\{y>0\}}$, then (8) can be written as:

$$(r(y) + \rho(y))\, f(y) = \text{net } b - \text{inflow for } y > 0.$$

Thus the original system with abandonment and/or impatient customers has the same probability law as an Abandonment-Only system with abandonment rates $k_a' \triangleq k_a + \kappa_a \leq \bar{\bar{k}}_a$ and $k_b' \triangleq k_b + \kappa_b \leq \bar{\bar{k}}_b$.

Comparing the original system to the Abandonment-Only and the Impatient-Only systems we have

PROPOSITION 2. *If* $-k_a - \lambda_b^I \omega_b^I < \lambda_b \omega_b - \lambda_a \omega_a < \lambda_a^I \omega_a^I + k_b$, *then*

$$\hat{\bar{\mathbb{P}}}_0 < \mathbb{P}_0 < \bar{\bar{\mathbb{P}}}_0.$$

The proposition implies that the Abandonment-Only system with abandonment rates $k_i = \zeta_i$, $i \in \{a, b\}$ is more likely to be empty than the Impatient-Only system with a total impatient arrival rate given by $\lambda_j^I \omega_j^I = \zeta_i$, $j \neq i$. The difference can be explained by observing that both systems have the same total flow and outflow rate of patient customers served by other patient customers. Therefore they have the same outflow of customers that are not served pairs of patient customers. However, the outflow in the Abandonment-Only system has lower variability than the outflow in the Impatient-Only system. Because of lower variability, the system has less queueing and therefore a higher probability of being empty.

4.2. System Performance

In this subsection we present the expectations of several metrics of interest in a double-sided queue, in particular the average system time and the system fill rate. These results are useful in analyzing the performance of our model in Section 6.

4.2.1. Mean Queue Length Let Q_a be the expected conditional number of side-a units in the system given $y<0$. By (7)

$$Q_a \triangleq \mathbb{E}[Y|Y<0] = \frac{\int_{-\infty}^{0}(-y)f(y)\,dy}{\int_{-\infty}^{0}f(y)\,dy} = \frac{U_a\int_{-\infty}^{0}(-y)e^{\Theta_a y}\,dy}{U_a\int_{-\infty}^{0}e^{\Theta_a y}\,dy} = \frac{1}{\Theta_a}.$$

Similarly,

$$Q_b \triangleq \mathbb{E}[Y|Y>0] = \frac{1}{\Theta_b}.$$

Let Q_{ABS} be the absolute queue length. Noting from Theorem 1, $P(Y<0) = U_a/\Theta_a$ and $P(Y>0) = U_b/\Theta_b$, so that

$$Q_{ABS} \triangleq E[|Y|] = Q_a\cdot\mathbb{P}(Y<0) + Q_b\cdot\mathbb{P}(Y>0) = \frac{U_a}{\Theta_a^2} + \frac{U_b}{\Theta_b^2}.$$

The expected queue length Q_{EXP} is

$$Q_{EXP} \triangleq \mathbb{E}[Y] = \int_{-\infty}^{\infty} yf(y)\,dy = \frac{U_a}{\Theta_a^2} - \frac{U_b}{\Theta_b^2}.$$

Q_a and Q_b measure the expected excess on each side. Q_{EXP} measures the bias of one side versus the other. Q_{ABS} measures the total excess in the system. When considering the application of our model to a financial market, this would be an expression of the liquidity of the market.

4.2.2. Average System Time Let W_i, $i=\{a,b\}$, be the average time a patient side-i unit spends in the system. By Little's law,

$$W_a = \frac{E[Y\cdot 1_{Y<0}]}{\lambda_i\omega_i} = \frac{E[Y|Y<0]P(Y<0)}{\lambda_a\omega_a} = \frac{U_a}{\lambda_a\omega_a\Theta_a^2} \qquad \text{and} \qquad W_b = \frac{U_b}{\lambda_b\omega_b\Theta_b^2}. \tag{9}$$

This measures the time in the system irrespective of whether the units are served by (transact with) orders from the opposite side or abandon the system. Impatient customers have a system time of exactly zero.

4.2.3. Fill Rate Units arriving to the system can exit either by transacting with counterparties or by abandoning the system. Let the fill rate, ϕ_i, $i=\{a,b\}$ be the fraction of side-i customers that transact. The fill rate equals 1 minus the fraction of units that abandon or depart on arrival. Consider side-a. The total arrival rate is $\lambda_a\omega_a + \lambda_a^I\omega_a^I$. The rate at which side-a customers abandon the queue is $k_aP(Y<0) = k_aU_a/\Theta_a$. Side-$a$ impatient customers do not transact if the queue length $y<x$ where x is the batch size of the arriving impatient customer. If $y\le 0$ then all arriving units in x do not transact. If $0<y<x$, then $x-y$ units do not transact. Recalling the expected number

of units per side-a impatient customer is ω_a^I, the expected number of impatient side-a units that do not transact per arrival is given by

$$\begin{aligned}\omega_a^I P(Y \le 0) + \int_{y=0}^{\infty}\int_{x=y}^{\infty}(x-y)\mu_a^I e^{-\mu_a^I x} U_b e^{-\Theta_b y} dx dy \\ = \omega_a^I \left(1 - \frac{U_b}{\Theta_b} + \frac{U_b}{\Theta_b + \mu_a^I}\right) \\ = \omega_a^I \left(1 - \frac{U_b}{\Theta_b} + \frac{\omega_a^I U_b}{\omega_a^I \Theta_b + 1}\right).\end{aligned}$$

Thus the expected rate of impatient side-a units that depart unserved per unit time is

$$\lambda_a^I \omega_a^I \left(1 - \frac{U_b}{\Theta_b} + \frac{\omega_a^I U_b}{\omega_a^I \Theta_b + 1}\right).$$

Then the fill rate is 1 minus the fraction of units that exit through abandonment or impatience. Generalizing for both sides, we have

$$\phi_i \triangleq 1 - \frac{\frac{k_i U_i}{\Theta_i} + \lambda_i^I \omega_i^I \left(1 - \frac{U_j}{\Theta_j} + \frac{\omega_i^I U_j}{\omega_i^I \Theta_j + 1}\right)}{\lambda_i \omega_i + \lambda_i^I \omega_i^I} \text{ for } i \in \{a, b\}, j \ne i. \tag{10}$$

4.3. Impact of Flow Characteristics on System Performance

Our closed-form expressions provide useful tools to study the impact of flow characteristics on system performance. We consider how the system changes with (1) asymmetric demand; (2) varying abandonment; and (3) increasing burstiness of demand on one side of the queue. To simplify, we let $\lambda_a^I = \lambda_b^I = 0$. Figure 2 shows the expected fill rate and system times for side-a and side-b customers, and their weighted average, while varying λ_a and λ_b between 0.8 and 1.2, holding $\lambda_a + \lambda_b = 2$ (here we let $\omega_a = \omega_b = 100$ and $k_a = k_b = 37.5$). We observe, as is intuitive, that the fill rate decreases and the system time increases for side-a customers as their prevalence increases. Further, the weighted average fill rate given by $\phi_{\text{avg}} = \frac{\lambda_a}{\lambda_a + \lambda_b}\phi_a + \frac{\lambda_b}{\lambda_a + \lambda_b}\phi_b$ is maximized at symmetry ($\lambda_a = \lambda_b = 1$), though it is somewhat insensitive to changing parameter values. A nearly identical figure holds for the case where ω_a and ω_b vary from 80 to 120, holding $\omega_a + \omega_b = 200$, and $\lambda_a = \lambda_b = 1$, $k_a = k_b = 37.5$. This demonstrates that the system behavior is closely related to the overall arrival rate for each side, i.e., $\lambda_i \omega_i$, $i \in \{a, b\}$.

Table 1 shows the fill rate and system time for both side-a and side-b customers for several cases. We observe increasing the abandonment rate reduces both the fill rates and system times. For the asymmetric case, there are a total of $\lambda_a \omega_a = 100$ units per unit time arriving on side-a while $\lambda_b \omega_b = 75$. Again we observe the side with the higher arrival rate has lower fill rates and higher system times.

Next we consider the trade-off between the fill rate and system time when both sides bring the same number of units to the system per unit time, but one side brings either few, large orders, and

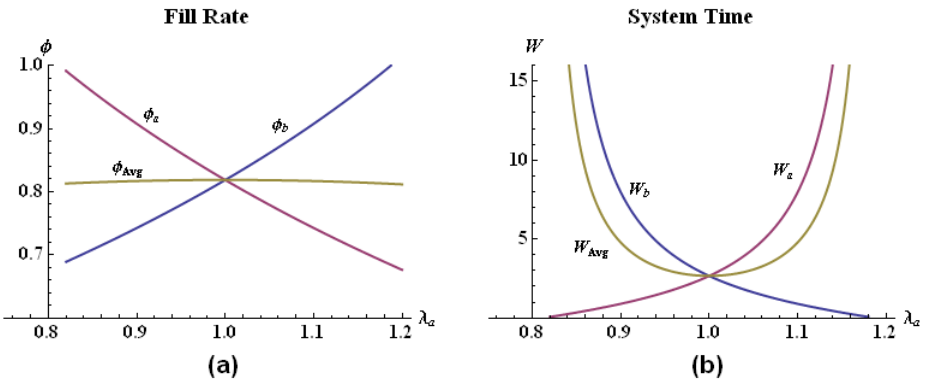

Figure 2 Expected fill rates/system times for side-a and b customers and their weighted average, varying λ_a & λ_b.

Table 1 Fill rate and system times for both side-a and side-b customers

System Level Metric	**Symmetric**	$\lambda_a = \lambda_b = 1$, $\omega_a = \omega_b = 100$	**Asymmetric**	$\lambda_a = 1$, $\lambda_b = 5$, $\omega_a = 100$, $\omega_b = 15$
	Low Abandonment $k_a = k_b = 37.5$	High Abandonment $k_a = k_b = 1000$	Low Abandonment $k_a = k_b = 37.5$	High Abandonment $k_a = k_b = 1000$
$\phi_\mathbf{a}$	81.87%	9.50%	69.03%	7.08%
$\phi_\mathbf{b}$	81.87%	9.50%	92.04%	9.44%
$\mathbf{W_a}$	2.67	0.10	8.09	0.11
$\mathbf{W_b}$	2.67	0.10	0.22	0.02

the other side brings more small orders. To this end, we let $\lambda_a = 1$, $\omega_a = 100$, $k_a = k_b = 37.5$. Here, $\lambda_b \in [0, 1.2]$ and we vary λ_b and ω_b such that $\lambda_b \omega_b = 100$. In Figure 3, we parametrically plot the fill rate and system time. For both curves the system time increases as λ_b decreases implying that to the left (right) of the peak the side-b arrival rate is higher (lower) than the side-a arrival rate. We observe that the side that has more frequent arrivals has a lower system time. Noting that the maximum fill rate occurs at symmetry and that by definition both sides have the same fill rate, we observe that the side with more frequent, smaller orders performs better, i.e., higher fill rate and lower system time.

5. Customer-Level Measures

In this section, we consider the experience of an individual customer who brings a known number of units to the system and who has a service deadline, i.e., a given maximum time they are willing to wait in queue. We assume this customer (referred to as the marked customer) arrives to the steady-state queue modeled in Section 3. We study the expected system time and fill rate for such a customer noting that a marked customer's experience will differ from the expectations presented in the previous section because of these specific attributes.

We consider two types of marked customers, *time-constrained* and *time-unconstrained*. A time-constrained individual abandons the queue at deadline T, withdrawing all of his unserved units at

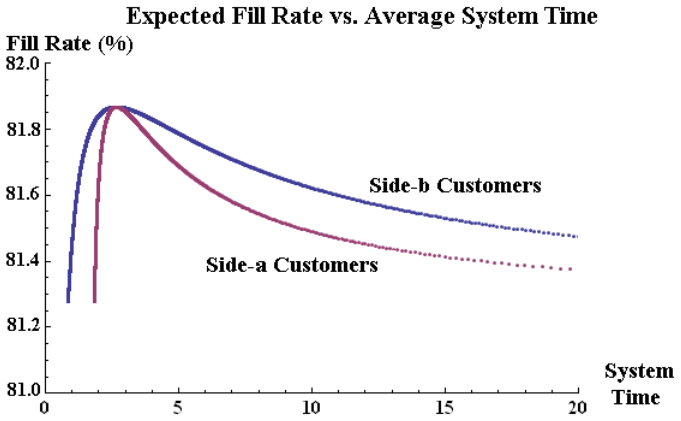

Figure 3 A parametric plot of the expected fill rate versus the average system time of side-a & side-b customer.

that time. None of his units abandon the system prior to this time. This is a common assumption in modeling abandonment from a queue. For such a customer, we present his expected fill rate. Alternatively, a time-unconstrained individual stays in the queue until her entire order is processed. That is, she is of infinite patience. By definition, her fill rate will be 100% and we derive her expected system time.

5.1. Clearing Time

Consider a side-b marked customer and suppose there are side-b units in queue upon his arrival (i.e., $y > 0$). Assume that none of these units abandon the queue, so that their departure is triggered exclusively by the arrival of both patient and impatient side-a customers. Define the clearing time $\tau_b^c(y)$ as the time for these side-b units to leave the system by being served. We will use this clearing time in computing the expected system time of the marked customer.

Let $\bar{Z}_a(t)$ be the number of side-a units that arrive to the system in time t assuming $\bar{Z}_a(0) = 0$. That is,

$$\bar{Z}_a(t) \triangleq \sum_{i=1}^{N_a(t)} X_{a,i} + \sum_{i=1}^{N_a^I(t)} X_{a,i}^I,$$

with $\{N_a(t), t \geq 0\}$ and $\{N_a^I(t), t \geq 0\}$ and $X_{a,i}$ and $X_{a,i}^I$ as defined in Section 3. Let $\bar{N}_a(t) = N_a(t) + N_a^I(t)$. Then $\{\bar{N}_a(t), t \geq 0\}$ is a Poisson process with rate $\bar{\lambda}_a = \lambda_a + \lambda_a^I$. Let $\bar{X}_{a,i}$ be a random variable with density $g_a(z)$ and distribution function $G_a(z)$ given by

$$g_a(z) = \frac{\lambda_a \mu_a e^{-\mu_a z} + \lambda_a^I \mu_a^I e^{-\mu_a^I z}}{\bar{\lambda}_a}, \qquad G_a(z) = 1 - \frac{\lambda_a e^{-\mu_a z} + \lambda_a^I e^{-\mu_a^I z}}{\bar{\lambda}_a}, \tag{11}$$

respectively. Then $\bar{Z}_a(t)$ is the compound Poisson process

$$\bar{Z}_a(t) = \sum_{i=1}^{\bar{N}_a(t)} \bar{X}_{a,i}.$$

The clearing time for y units is given by

$$\tau_b^c(y) \triangleq \inf\left\{t \geq 0 : \bar{Z}_a(t) > y\right\}.$$

PROPOSITION 3. *The expected value of $\tau_b^c(y)$ is*

$$\mathbb{E}[\tau_b^c(y)] = \frac{\lambda_a\omega_a^2 + \lambda_a^I\left(\omega_a^I\right)^2}{\left(\lambda_a\omega_a + \lambda_a^I\omega_a^I\right)^2} + \frac{y}{\lambda_a\omega_a + \lambda_a^I\omega_a^I} - \frac{\lambda_a\lambda_a^I\left(\omega_a - \omega_a^I\right)^2}{\left(\lambda_a + \lambda_a^I\right)\left(\lambda_a\omega_a + \lambda_a^I\omega_a^I\right)^2} e^{-y\frac{\left(\lambda_a\omega_a + \lambda_a^I\omega_a^I\right)}{\left(\lambda_a + \lambda_a^I\right)\omega_a\omega_a^I}}.$$

A similar expression holds for the clearing time for the side-a queue.

5.2. System Time for a Time-Unconstrained Customer

Consider a side-i, time-unconstrained marked customer who submits an order of size x and remains in the system until her entire order is filled. Let $T_i^c(x,y)$ be her system time given the queue length is y upon her arrival. We assume any side-i units in queue upon her arrival abandon the system at rate k_i. Note this assumption is consistent with a first come/first abandon assumption which was not made in Section 3 but is needed here. Let $\mathbb{E}[T_i^c(x,y)]$, $i = \{a,b\}$ be her conditional expectation system time and let $\mathbb{E}[T_i^c(x)]$ be her unconditional expected system time.

Suppose she is a side-b arrival. There are three cases: (1) If $y \leq -x$, all her units are served on arrival and $T_b^c(x,y) = 0$. (2) If $-x < y \leq 0$, then $x+y$ of her units remain in queue and $\mathbb{E}[T_b^c(x,y)] = \mathbb{E}[\tau_b^c(x+y)]$, the time to clear these units. (3) If $y > 0$, she must queue and $\mathbb{E}[T_b^c(x,y)]$ solves the following integro-differential equation obtained by conditioning on the first arrival of a side-a customer. Suppose the first arrival to side-a brings z units. Then

$$\begin{aligned}\mathbb{E}[T_b^c(x,y)] = &\int_{\frac{y}{k_b}}^{\infty} \bar{\lambda}_a e^{-\bar{\lambda}_a t}\left(t + \int_0^x g_a(z)\mathbb{E}[\tau_b^c(x-z)]\,dz\right)dt \\ &+ \int_0^{\frac{y}{k_b}} \bar{\lambda}_a e^{-\bar{\lambda}_a t}\left(t + \int_0^{y-tk_b} g_a(z)\mathbb{E}[T_b^c(x, y-tk_b-z)]\,dz \right. \\ &\qquad\qquad \left. + \int_{y-tk_b}^{x+y-tk_b} g_a(z)\mathbb{E}[\tau_b^c(x+y-tk_b-z)]\,dz\right)dt. \end{aligned} \tag{12}$$

The first term expresses the case when the first arrival from side-a occurs at time t after all y units in queue abandon. Her remaining system time is $\mathbb{E}[\tau_b^c(x-z)]$. The second term expresses a renewal equation for the case where the first arrival occurs at t before all y units have abandoned. In that case if $z < y - tk_b$, then $y - tk_b - z$ units remain in queue in front of her. Her expected remaining system time is $\mathbb{E}[T_b^c(x, y-tk_b-z)]$. If $z \geq y - tk_b$, then $x + y - tk_b - z$ of her units remain and her additional system time is $\mathbb{E}[\tau_b^c(x+y-tk_b-z)]$. The time until the first arrival is negatively exponentially distributed with mean $1/\bar{\lambda}_a$ and the order size is distributed as $g_a(z)$.

PROPOSITION 4. *The conditional expected system time given that a time-unconstrained side-b customer submits an order of size x to a queue that contains y units is*

$$\mathbb{E}[T_b^c(x,y)] = \begin{cases} 0 & -\infty < y \le -x, \\ \mathbb{E}[\tau_b^c(x+y)] & -x < y \le 0, \\ \mathbb{E}\left[\bar{T}_b^c(x,y)\right] & 0 < y < \infty. \end{cases}$$

where

$$\begin{aligned}\mathbb{E}\left[\bar{T}_b^c(x,y)\right] =& C_0 + C_1 y + \frac{1}{2} C_1 \frac{K_1 + \sqrt{K_2}}{K_1\sqrt{K_2} - K_2} e^{-y\left(K_1 - \sqrt{K_2}\right)} - \frac{1}{2} C_1 \frac{K_1 - \sqrt{K_2}}{K_1\sqrt{K_2} + K_2} e^{-y\left(K_1 + \sqrt{K_2}\right)} \\ &+ \frac{x}{\lambda_a\omega_a + \lambda_a^I\omega_a^I} - \frac{\lambda_a\lambda_a^I\left(\omega_a - \omega_a^I\right)^2}{\left(\lambda_a + \lambda_a^I\right)\left(\lambda_a\omega_a + \lambda_a^I\omega_a^I\right)^2} e^{-x\frac{\left(\lambda_a\omega_a + \lambda_a^I\omega_a^I\right)}{\left(\lambda_a + \lambda_a^I\right)\omega_a\omega_a^I}}.\end{aligned}$$

with

$$\begin{aligned} K_1 &= \left(\lambda_a + \lambda_a^I + k_b\mu_a + k_b\mu_a^I\right)/2k_b, \\ K_2 &= \left(\lambda_a^2 + (\lambda_a^I)^2 + k_b^2\left(\mu_a - \mu_a^I\right)^2 - 2k_b\lambda_a^I\left(\mu_a + 3\mu_a^I\right) + 2\lambda_a\left(\lambda_a^I + k_b\left(\mu_a - \mu_a^I\right)\right)\right)/4k_b^2, \\ C_0 &= \frac{\left(K_1^2 - K_2\right)\left(\lambda_a^I\left(\omega_a^I\right)^2 + \lambda_a\omega_a^2\right)\left(k_b + \lambda_a\omega_a + \lambda_a^I\omega_a^I + 2\lambda_a^I\omega_a\right) - 2K_1\left(\lambda_a^I\omega_a^I + \lambda_a\omega_a\right)^2}{\left(K_1^2 - K_2\right)\left(k_b + \lambda_a\omega_a + \lambda_a^I\omega_a^I + 2\lambda_a^I\omega_a\right)\left(\lambda_a^I\omega_a^I + \lambda_a\omega_a\right)^2}, \\ C_1 &= \left(k_b + \lambda_a\omega_a + \lambda_a^I\omega_a^I + 2\lambda_a^I\omega_a\right)^{-1}. \end{aligned}$$

Observe that when there is a side-b queue, the system time has a similar structure to the clearing time, having both linear and exponential terms in x and y. The expected system time for a time-unconstrained side-b individual with an order of size x is given by

$$\mathbb{E}[T_b^c(x)] = \int_{-x}^{\infty} \mathbb{E}[T_b^c(x,y)]\, dF(y). \tag{13}$$

Since the queue length density function from (7) is of exponential type, $\mathbb{E}[T_b^c(x)]$ can be solved as a closed-form expression. Similar expressions hold for $\mathbb{E}[T_a^c(x,y)]$ and $\mathbb{E}[T_a^c(x)]$.

5.3. Fill Rate for a Time-Constrained Customer

We now consider a side-i, time-constrained marked customer who submits an order of size x and remains in the system until his order is filled or his deadline, T time units, have passed. We assume, without loss of generality, that he enters into the system in steady-state at time 0. None of his units abandon until time T, when all his remaining unserved units abandon the system en masse. As in Section 5.2, we assume units in queue in front of him abandon the system at rate k_i. Let $\phi_i^c(x,T)$ be the customer's fill rate, i.e., the expected fraction of the x units that are served.

Suppose he is a side-b arrival. To find $\phi_b^c(x,T)$ we consider several cases. First, if $T=0$, the customer is an "impatient" customer and directly we know

$$\begin{aligned}\phi_b^c(x,0) &= \frac{E[\min(x,\max[-Y,0])]}{x} \\ &= \frac{U_a\left(1-e^{-x\Theta_a}\right)}{x\Theta_a^2}\end{aligned} \tag{14}$$

For $T>0$, we find $\phi_b^c(x,T)$ by conditioning on Y, the queue length at the arrival time of the marked customer, τ, the time when only units from the marked customer remain in queue, $\bar{Z}_a(\tau)$, the number of counter-party units that arrive by time τ, and $\bar{Z}_a(\tau,T)$, the number of counter-party units that subsequently arrive prior to T. We find $\phi_b^c(x,T|Y,\tau,\bar{Z}_a(\tau),\bar{Z}_a(\tau,T))$, the conditional fill rate. We then derive the probability distribution for each conditioning event and then de-condition to find $\phi_b^c(x,T)$.

Let y be the realization of Y. Consider the case where $y>0$, i.e., there is a side-b queue in front of the marked customer and units in queue ahead of the marked customer abandon the queue at rate k_b. Then τ is given by the time that the process $\bar{Z}_a(t)$ either hits or jumps over the linearly decreasing boundary given by $y-k_bt$. That is τ is a stopping time given as

$$\tau = \inf\{t\geq 0|\bar{Z}_a(t)\geq y-k_bt\}.$$

Observe if $y\leq 0$ then $\tau=0$. Otherwise, $\tau>0$.

If $y>0$, the fill rate depends on whether the last of the queued units in front of the marked customer abandons or is served. That is, for $y>0$ the fill rate depends on whether $\bar{Z}_a(t)$, respectively, hits or jumps over the linearly decreasing boundary $y-k_bt$. Let $Q=(\bar{Z}_a(\tau)-(y-k_b\tau))^+$ for $y>0$. Q is the overshoot. If $\bar{Z}_a(t)$ hits the boundary, $Q=0$. Otherwise, $Q>0$ (almost surely). So given y and τ, conditioning on Q is equivalent to conditioning on $\bar{Z}_a(\tau)$. Let q be a realization of Q. Let z be the realization of $\bar{Z}_a(\tau,T)$, the number of units that arrive between τ and T.

The conditional fill rate is then defined as $\phi_b^c(x,T|y,\tau,q,z)$ and is given by

$$\phi_b^c(x,T|y,\tau,q,z) = \begin{cases} \frac{z-y}{x}\bigwedge 1 & \text{if } y\leq 0 \\ \frac{z+q}{x}\bigwedge 1 & \text{if } y>0,\tau\leq T, \\ 0 & \text{if } y>0,\tau>T. \end{cases} \tag{15}$$

(15) expresses the following reasoning: For $y\leq 0$, $-y$ units are served immediately and the remainder are served by the arrivals z. If $y\leq -x<0$ or if $z>x+y$, $\phi_b^c(x,T|y,\tau,z,q)=1$; otherwise $0\leq z<x+y$ and $\phi_b^c(x,T|y,\tau,z,q)=(z-y)/x$. If $y>0$ and $\tau\leq T$, q units are served by the overshoot and the remainder are served by the arrivals z. If $y>0$ and $\tau>T$, $\bar{Z}_a(\tau,T)=0$ and $\phi_b^c(x,T|y,\tau,q,z)=0$.

In order to de-condition $\phi_b^c(x,T|y,\tau,q,z)$ we require the probability distributions for $\bar{Z}_a(\tau,T)$ and Q. Because the arrivals are Poisson, $\bar{Z}_a(\tau,T) \sim \bar{Z}_a(T-\tau)$ for $\tau \leq T$. Let $\psi_t(z)$ be the distribution of $\bar{Z}_a(t)$. Recalling the jump size distribution $g_a(z)$ associated with $\bar{Z}_a(t)$ defined in (11), we have

$$\psi_t(z) = \sum_{n=1}^{\infty} e^{-\bar{\lambda}_a t} \frac{\left(\bar{\lambda}_a t\right)^n}{n!} g_a^{*n}(z), \tag{16}$$

where $g_a^{*n}(z)$ is the n-fold convolution of $g_a(z)$. Example: If $\lambda_a^I = 0$ then (16) becomes

$$\psi_t(z) = e^{-\lambda_a t} e^{-\mu_a z} \sum_{n=1}^{\infty} \frac{(\lambda_a t)^n}{n!} \frac{\mu_a^n z^{n-1}}{(n-1)!} = e^{-\lambda_a t} e^{-\mu_a z} \sqrt{\frac{\lambda_a \mu_a t}{z}} I_1\left(z\sqrt{\lambda_a \mu_a z t}\right)$$

where $I_1(\cdot)$ is the modified Bessel function of the first order.

To determine the distribution of Q we need the probability that it is zero or positive given $y > 0$. To do so we define the defective densities for the time the process alternately hits $\{h(t,y)\}$ or jumps over $\{j(t,y)\}$ the boundary $y - k_b t$:

$$h(t,y) = \frac{d}{dt}\mathbb{P}(\tau \leq t, \bar{Z}_a(t) = y - k_b t),$$
$$j(t,y) = \frac{d}{dt}\mathbb{P}(\tau \leq t, \bar{Z}_a(t) > y - k_b t).$$

Following Perry et al. (1999) and Zacks (2004),

$$h(t,y) = e^{-\lambda t} \sum_{n=1}^{\infty} \frac{(\lambda t)^n}{n!} g_a^{*(n)}(y - k_b t),$$
$$j(t,y) = \lambda e^{-\lambda t} \left(\frac{\lambda_a e^{-\mu_a (y-k_b t)} + \lambda_a^I e^{-\mu_a^I (y-k_b t)}}{\lambda_a + \lambda_a^I} \right) + e^{-\lambda t} \sum_{n=1}^{\infty} \frac{(\lambda t)^n}{n!} \left(G_a^{*(n)}(y - k_b t) - G_a^{*(n+1)}(y - k_b t) \right),$$

where $g_a^{*(n)}(y - k_b t)$ and $G_a^{*(n+1)}(y - k_b t)$ are the n-fold convolutions of $g_a(y - k_b t)$ and $G_a(y - k_b t)$, respectively.

Example: If $\lambda_a^I = 0$ then

$$\begin{aligned} h(t,y) &= k_b \mu_a e^{-\lambda_a t} e^{-\mu_a (y-k_b t)} \sum_{n=0}^{\infty} \frac{(\lambda_a t)^{n+1}}{(n+1)!} \frac{\mu_a^n (y - k_b t)^n}{n!} \\ &= k_b e^{-\lambda_a t} e^{-\mu_a (y-k_b t)} \sqrt{\frac{\lambda_a \mu_a t}{y - k_b t}} I_1\left(2\sqrt{\lambda_a \mu_a t (y - k_b t)}\right), \text{ and} \\ j(t,y) &= \lambda_a e^{-\lambda_a t} e^{-\mu_a (y-k_b t)} \sum_{n=0}^{\infty} \frac{(\lambda_a t)^n}{n!} \frac{\mu_a^n (y - k_b t)^n}{n!} \\ &= \lambda_a e^{-\lambda_a t} e^{-\mu_a (y-k_b t)} I_0\left(2\sqrt{\lambda_a \mu_a t (y - k_b t)}\right) \end{aligned}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel function of zeroth and first order, respectively.

If $\tau < y/k_b$ the hitting time has density $h(\tau,y)$ while the jump time has density $j(\tau,y)$. If $\tau = y/k_b$, the process (almost surely) hits the boundary at $\bar{Z}_a(y/k_b) = 0$ (i.e., there were no arrivals before

time y/k_b). This occurs with probability $\mathbb{P}(\tau = y/k_b) = \text{Exp}[-y\lambda_a/k_b]$. In either the hit or jump over case, if $\tau > T$, $\phi_b^c(x, T|y, z) = 0$.

From (15) and taking the expectation over y, the expected fill rate $\phi_b^c(x, T)$ is

$$\begin{aligned}\phi_b^c(x,T) = &\int_{y-\infty}^{-x} dF(y) + \int_{-x}^{0}\int_{z=0}^{\infty}\left(\frac{z-y}{x}\bigwedge 1\right)\psi_{T-t}(z)\,dz\,dF(y)\\ &+\int_{y=0^+}^{\infty}\int_{t=0}^{\min(T,\frac{y}{k_b})}\int_{z=0}^{\infty}\left(\frac{z}{x}\bigwedge 1\right)\psi_{T-t}(z)h(t,y)\,dz\,dt\,dF(y)\\ &+\int_{y=0^+}^{k_bT}\int_{z=0}^{\infty}\left(\frac{z}{x}\bigwedge 1\right)e^{-y\frac{\lambda_a}{k_b}}\psi_{T-\frac{y}{k_b}}(z)\,dz\,dF(y)\\ &+\int_{y=0^+}^{\infty}\int_{t=0}^{\min(T,\frac{y}{k_b})}\int_{q=0}^{\infty}\int_{z=0}^{\infty}\left(\frac{z+q}{x}\bigwedge 1\right)\psi_{T-t}(z)g_a(q)j(t,y)\,dz\,dq\,dt\,dF(y).\end{aligned} \tag{17}$$

A similar formula holds for a side-a individual.

6. Numerical Results

In this section, we compare the results of our model based on a constant abandonment rate to a simulated system in which abandonment is at the customer level. That is, customers arrive according to a Poisson process, bring an exponentially distributed number of units in their order, and abandon the system at an exponentially distributed deadline time, taking all of their unserved units from the system at that time. We consider how closely the key performance measures given by our model, i.e., the expected fill rate and system time, match those of the simulated system.

We present numerical results for the models at the system- and customer-level. We address three questions: (1) What is the general behavior of these metrics and what are the sensitivities to changing parameter values? (2) How do the system-level metrics compare with results from the simulation? (3) How do the customer-level metrics compare with those given by this simulation?

To summarize our results, we find that the measures we derive above are good in that they provide insight into the system behavior while comparing favorably with the simulation. We show that high service levels are achieved with frequent, small arrivals, that the measures are more sensitive to changes in the demand rate, and less sensitive to changes in the order size. For both system- and customer-level metrics there is a close correspondence between the model and the simulated system. We conclude our model is quite attractive, both because of its simplicity and tractability, and because of its accuracy in approximating the performance in the simulated system.

6.1. System-Level Comparison to a Simulated Model with Deadlines

In the simulation we choose parameter values for λ_i and ω_i, and let $\lambda_i^I = 0$, $i = a, b$. We let t_i, $i = a, b$, be the mean deadline time for simulated customers arriving to side i. For each test case we simulate 5,000,000 customer arrivals. We find the average time a unit spends in the system

(the observed system time) and the percentage of units served (the observed fill rate). In order to determine the expected fill rate and system time using the results in (9) and (10), we require the values of k_i, $i = a, b$. We use the observed abandonment rates, k_i, given by the number of units that abandon the system from side $i = a, b$ divided by the time the system spent on side i.

In Table 2, we present the fill rate and system time for side-b that we observe in the simulation and the expected values given by our model. We also present the relative difference measured by (Observed-Predicted)/Predicted$\times$100%. We observe the fill rate and system time are both increasing in the customers' patience (increasing $t_a = t_b$). The system is sensitive to changes in arrival rate (increasing $\lambda_a = \lambda_b$) and relatively insensitive to changes in the order size (increasing $\omega_a = \omega_b$), when demand is balanced. However, increasing the frequency of orders while decreasing the order size to hold the total demand constant (changing $\lambda_a = \lambda_b$ while $\lambda_a\omega_a = \lambda_b\omega_b = 100$), increases the fill rate and decreases the system time. This indicates that one may expect, for any level of customer patience, a greater number of smaller orders. For asymmetric systems we observe increased demand from the opposite side increases the fill rate and lowers the system time. With respect to the model's agreement with the simulation, we observe a very close correspondence between the expected and observed fill rates. This is to be expected. We are specifically comparing our system for an assumed abandonment rate with a simulation where the assumed average patience of the customers results in this abandonment rate. However, because simulated customers depart the system with all of their units at the same time, the actual fill rate can differ from our expectations.

It is the difference in the observed and predicted system times that measures the mis-specification of the model. We observe that the accuracy of the system time depends on the mean patience of the simulated customers and the mean arrival rate of the customers. Specifically, we observe that if the mean deadline is larger than the mean inter-arrival time ($t_a = t_b > 1/\lambda_a = 1/\lambda_b$), the customers are likely to queue, resulting in higher fill rates. For these cases we observe a relative difference of approximately 30% between the observed and predicted system times. However, when $t_a = t_b < 1/\lambda_a = 1/\lambda_b$, customers are relatively impatient and we observe higher accuracy in the predictions. For these latter cases, almost all customers abandon the system. In the model, the expected system time (for side-b, say) would be determined by linear abandonment as ω_b/k_b. That is, the equivalent mean deadline t_b would be ω_b/k_b. Indeed, we observe that as the mean deadline of the simulated customers decreases, it approaches ω_i/k_i resulting in higher accuracy in the system time in cases with lower fill rates.

6.2. Customer-Level Comparison to a Simulated Model with Deadlines

We now consider how the customer level measures in our model compare with the performance of an individual customer in the simulated system. To do so, we simulate the system as before but

Table 2 Observed Rate of Abandonment (k_a, k_b), and Observed and Predicted System-level Fill Rate and System Time for side-**b**. The parameter values are $\lambda_a = 1.0$, $\lambda_\mathbf{b} = \mathbf{1.0}$, $\omega_\mathbf{a} = \mathbf{100.0}$, $\omega_\mathbf{b} = \mathbf{100.0}$, $\mathbf{t_a = 0.10}$, $\mathbf{t_b = 0.10}$ except for the varied value(s) as noted in the first two columns of the table.

		Rate of Abandonment		**Fill Rate (%)**			**System Time**		
Varying Param.	**Value**	k_a	k_b	**Obs.**	**Pred.**	**Rel Diff(%)**	**Obs.**	**Pred.**	**Rel Diff(%)**
$t_a = t_b$	0.01	10062	10062	0.98	0.99	-0.71	0.01	0.01	-0.00
	0.05	2053	2053	4.77	4.82	-1.06	0.05	0.05	-2.42
	0.10	1052	1052	8.95	9.14	-2.08	0.09	0.10	-4.14
	0.50	248.3	248.3	30.9	32.5	-4.64	0.35	0.40	-14.2
	1.00	145.9	145.9	45.1	47.3	-4.70	0.55	0.69	-19.8
	5.00	54.3	54.3	73.6	74.6	-1.30	1.32	1.84	-28.3
	10.0	37.5	37.5	81.4	81.9	-0.50	1.85	2.67	-30.6
	20.0	26.3	26.3	87.0	87.1	-0.10	2.61	3.80	-31.4
	100.0	11.6	11.6	94.2	94.3	-0.01	5.68	8.62	-34.2
$\lambda_a = \lambda_b$	0.10	1005	1005	0.99	0.99	-0.20	0.10	0.10	-0.51
	0.50	1025	1025	4.71	4.76	-1.15	0.10	0.10	-2.37
	5.00	1240	1240	30.9	32.5	-4.86	0.07	0.08	-14.3
	10.00	1458	1458	45.1	47.3	-4.78	0.05	0.07	-19.9
$\omega_a = \omega_b$	10.00	105.0	105.0	8.90	9.08	-1.89	0.09	0.10	-4.35
	50.00	524.6	524.6	8.89	9.08	-2.04	0.09	0.10	-4.38
	200.00	2100	2100	8.89	9.07	-2.00	0.09	0.10	-4.32
	1000.00	10502	10502	8.91	9.07	-1.84	0.09	0.10	-4.44
$\lambda_a = \lambda_b$ (holding $\lambda_a\omega_a = \lambda_b\omega_b = 100$)	0.10	10046	10046	0.98	0.99	-0.73	0.10	0.10	-0.37
	0.20	5054	5054	1.95	1.96	-0.46	0.10	0.10	-0.94
	0.50	2050	2050	4.71	4.76	-1.11	0.10	0.10	-2.31
	2.00	549.1	549.1	16.0	16.6	-3.30	0.08	0.09	-7.78
	5.00	248.1	248.1	30.9	32.5	-4.85	0.07	0.08	-14.3
	10.00	145.9	145.9	45.1	47.3	-4.74	0.05	0.07	-20.0
	100.00	37.54	37.54	81.4	81.9	-0.56	0.02	0.03	-30.0
t_b	0.01	1050	10052	5.09	5.20	-2.13	0.01	0.01	-0.42
	0.05	1050	2050	6.82	6.97	-2.03	0.05	0.05	-2.26
	0.5	1049	247.7	21.5	22.5	-4.71	0.39	0.46	-15.3
	1.00	1049	145.6	31.5	33.3	-5.48	0.69	0.87	-21.1
	5.00	1050	54.34	59.1	60.3	-2.02	2.04	2.88	-29.0
	20.00	1049	26.13	77.2	77.5	-0.33	4.56	6.71	-32.1
λ_b	0.10	1050	1004	9.39	9.48	-0.98	0.09	0.09	-0.40
	0.50	1050	1025	9.13	9.29	-1.72	0.09	0.09	-2.09
	2.00	1049	1100	8.42	8.66	-2.76	0.09	0.10	-8.69
	10.00	1042	1567	5.79	6.19	-6.60	0.09	0.15	-38.6
λ_a	0.10	10078	1050	1.78	1.82	-2.09	0.10	0.10	-4.83
	0.50	2049	1050	6.15	6.27	-1.92	0.09	0.10	-4.56
	2.00	550.0	1048	11.4	11.8	-3.29	0.09	0.09	-4.02
	10.00	155.2	1048	14.8	17.4	-14.7	0.09	0.09	-1.45

randomly insert "marked" side-b customers with service deadline T. For each test case we simulate 10,000,000 customers arrivals with 100,000 inserted marked customers. Each such customer brings $x = \omega_b$, the average number of units for side-b customers. For time-constrained customers, we assume that T is either 0 or t_b, the mean deadline for a side-b customer. We find the simulated average fill rate and compare it with the expectation which we compute using (14) for $T = 0$ and (17)

for $T = t_b$. For a time-unconstrained marked customer, $T = \infty$ by definition. For time-unconstrained customers we find the simulated average system time and compare it to the expectation given in (13).

We make the following two key observations: (1) The results for the time-constrained and time-unconstrained customers indicate our model, and the approximations derived from it, are relatively accurate; (2) The table exhibits the basic trade-offs an individual customer makes in a double-sided queue.

We observe that in Table 3, the estimated fill rate for the time-constrained customers closely matches the observed fill rate. When $T = 0$, the expected fill rate is within 3% of the observed fill rate for all but two of the test cases. When $T > 0$, the expectation is always lower than the observed fill rate. The percentage error is largest when the absolute error is small. Similarly, we observe that the expected system time provides a reasonably accurate approximation of the observed customer system time, noting that when the fill rate is low, the system time approximation is very accurate.

We observe that the fill rate for $T = 0$ is approximately one-half that of the value at $T = t_b$ (except for the case where the marked customer's patience is very high). Comparing the case of $T = t_b$ to $T = \infty$, we observe the time required to fill an order completely may be many times longer than the time required to achieve a small percentage fill. We also observe that the expected time to achieve a 100% fill rate is inversely related to the arrival rate.

Table 3 Time-Constrained Fill Rate and Time-Unconstrained System Time. Marked customers arrive with $\mathbf{x}=\omega_b$ units. The parameter values are $\lambda_\mathbf{a}=\mathbf{1.0}$, $\lambda_\mathbf{b}=\mathbf{1.0}$, $\omega_\mathbf{a}=\mathbf{100.0}$, $\omega_\mathbf{b}=\mathbf{100.0}$, $\mathbf{t_a=0.10}$, $\mathbf{t_b=0.10}$ except for the varied value(s) as noted in the first two columns of the table.

		Time-Constrained Fill Rate (%), $T=0$			**Time-Constrained Fill Rate (%), $T=t_b$**			**Time-Unconstrained System Time ($T=\infty$)**		
Varying Param.	**Value**	**Obs.**	**Pred.**	**Rel Diff (%)**	**Obs.**	**Pred.**	**Rel Diff (%)**	**Obs.**	**Pred.**	**Rel Diff (%)**
$t_a=t_b$	0.01	0.62	0.62	0.0	1.24	0.99	25.0	1.98	1.98	0.00
	0.05	2.93	2.99	-2.0	5.96	4.81	23.9	1.96	1.96	0.00
	0.10	5.54	5.68	-2.5	11.3	9.26	22.0	1.91	1.91	-0.00
	0.50	19.3	19.8	-2.6	39.4	34.7	13.6	1.74	1.74	-0.00
	1.00	27.6	28.1	-1.6	57.1	51.9	10.1	1.71	1.72	-0.50
	5.00	41.9	41.3	1.5	90.0	86.1	4.52	2.05	2.32	-11.6
	10.0	44.8	44.2	1.4	97.2	93.1	4.44	2.52	3.00	-16.1
	20.0	46.4	46.1	0.6	99.7	97.1	2.72	3.26	4.04	-19.3
	100.0	47.6	48.4	-1.8	100.0	99.8	0.18	7.15	8.72	-18.0
$\lambda_a=\lambda_b$	0.10	0.63	0.63	0.79	1.24	0.99	25.3	19.9	19.9	0.04
	0.50	3.01	3.00	0.60	6.03	4.82	24.9	3.92	3.91	0.31
	5.00	19.0	19.8	-3.98	39.3	34.9	12.4	0.35	0.35	0.21
	10.00	27.5	28.1	-2.09	56.6	52.3	8.21	0.17	0.17	-1.28
$\omega_a=\omega_b$	10.00	5.61	5.69	-1.37	11.1	9.27	20.3	1.92	1.91	0.46
	50.00	5.60	5.69	-1.67	11.1	9.27	20.2	1.91	1.91	-0.18
	200.00	5.65	5.68	-0.54	11.1	9.27	20.1	1.91	1.91	0.08
	1000.00	5.64	5.69	-0.82	11.1	9.27	20.2	1.92	1.91	0.37
$\lambda_a=\lambda_b$ (holding $\lambda_a\omega_a=\lambda_b\omega_b=100$)	0.10	0.63	0.63	0.20	1.29	0.99	29.9	19.9	19.9	0.13
	0.20	1.24	1.24	0.35	2.46	1.97	25.0	9.87	9.90	-0.31
	0.50	2.94	3.00	-1.69	6.02	4.81	25.1	3.93	3.91	0.60
	2.00	9.97	10.3	-3.28	20.0	17.2	16.7	0.93	0.92	0.53
	5.00	19.1	19.8	-3.24	39.2	34.5	13.8	0.35	0.35	-0.03
	10.00	27.4	28.1	-2.47	57.1	50.8	12.3	0.17	0.17	-0.88
	100.00	44.9	44.2	1.47	97.1	88.6	9.64	0.03	0.03	-2.86
t_b	0.01	5.78	5.93	-2.60	6.35	4.25	49.4	1.91	1.90	0.28
	0.05	5.77	5.82	-0.90	8.63	6.56	31.6	1.91	1.91	0.11
	0.5	4.75	4.85	-1.92	28.3	25.6	10.4	1.99	2.01	-0.87
	1.00	4.22	4.17	1.00	42.8	38.7	10.6	2.15	2.20	-2.39
	5.00	2.52	2.49	1.47	84.3	78.6	7.31	3.23	3.67	-12.0
	20.00	1.36	1.41	-3.90	99.4	94.7	4.97	5.65	7.16	-21.1
λ_b	0.10	6.07	6.20	-2.22	12.0	9.80	22.5	1.91	1.90	0.27
	0.50	5.92	5.97	-0.82	11.6	9.56	21.8	1.90	1.91	-0.10
	2.00	5.03	5.17	-2.57	10.6	8.71	22.0	1.93	1.93	0.25
	10.00	2.54	2.37	6.99	7.25	5.24	38.3	2.00	2.03	-1.35
λ_a	0.10	0.92	0.93	-0.37	1.86	1.81	2.79	10.9	10.9	-0.44
	0.50	3.56	3.64	-2.29	7.27	6.49	12.0	2.90	2.91	-0.50
	2.00	7.37	7.82	-5.77	14.9	11.4	31.4	1.42	1.41	0.38
	10.00	8.54	12.3	-30.71	17.9	17.3	3.56	1.02	0.98	4.03

7. Concluding Remarks

We consider a double-sided, batch-arrival queue with abandonment. Under an assumption of a constant abandonment rate rather than batch abandonment, we derive the steady-state probability distribution for the queue. The system is stable if the abandonment plus the demand rate from impatient customers is sufficient to deplete any imbalance in the demand from patient customers. Further, we show the equivalence of the system behavior for cases without impatient customers and cases with only impatient customers. We derive the main system level measures of fill rate and system time. We also derive the expected performance measures for an individual with a known deadline and order size. For both the system- and customer-level measures, the numerical results compare the performance of our model to a simulation where customers depart after a deadline. We show that for cases of sufficient impatience, our model displays a high degree of fidelity to the simulated values. As they are tractable, they provide an attractive alternative to simulation.

The paper is motivated by the operation of crossing networks and we believe our model captures the main aspects of these trading venues. We can draw several implications for their behavior from the numerical studies conducted. We demonstrate that the side that brings more demand has a lower fill rate and higher system time. However, if we equate the total demand on both sides at a fixed value, we find the opposite is true. That is, the side with more frequent, smaller orders achieves a lower system time for the same fill rate. This is significant as it indicates that for traders seeking to have their orders filled quickly, there is incentive, in balanced systems, to continually attempt to place small orders. Further, we observe that for individual customers, when other customers have little patience (modeled through t_a and t_b), there is little incentive for themselves to be patient. Reducing their patience from the average to zero, results in a small absolute reduction in their fill rate. In comparison, the time required to have their orders completely filled is 20–200 times longer. Because traders are exposed to price risk while waiting, one would expect that they will be of limited patience, choosing to have only a small percentage of their orders filled. The implication is that for crossing networks, we can expect systems with high frequency, impatient customers bringing small orders.

The difficulty is that increasing impatience results in low trading volume, a central concern to crossing network operators. To alleviate this, crossing networks would have to provide incentives for customers to increase their patience. The anonymity of the market place does provide some incentive to be patient by allowing large traders to hide the true size of their position and in so doing, keep it in the market longer. Additional mechanisms include the allowance of limit orders, which protect patient users from wild price swings. As we note, there has been some research on the use of limit orders in visible markets. However, they have played a minor role in crossing networks to date, though future research should investigate their use in these markets.

References

Asmussen, S. (2003). *Applied probability and queues*. Springer.

Boxma, O.J., David, I., Perry, D., Stadje, W. (2011). A new look at organ transplantation models and double matching queues. *Probability in the Engineering and Informational Sciences*, **25**(2), 135-155.

Buti, S., Rindi, B., Werner, I. (2011). Dark pool trading strategies. *Charles A. Dice Center Working Paper*, No. 2010-6. Ohio State University.

Conolly, B. W., Parthasarathy, P. R., Selvaraju, N. (2002). Double-ended queues with impatience. *Computers and Operations Research*, **29**(14), 2053-2072.

Dobbie, J. M. (1961). A doubled-ended queuing problem of Kendall. *Operations Research*, **9**(5), 755-757.

Feng, R. (2008). A generalization of the discounted penalty function in ruin theory. *PhD Thesis in Actuarial Science*. Waterloo, Ontario, Canada.

Foucault, T., Kadan, O., Kandel, E. (2005). Limit order book as a market for liquidity. *Review of Financial Studies*, **18**(4), 1171-1217.

Gaur, K., Kashyap, B. (1973). The double-ended queue with limited waiting space. *Indian Journal of Pure Applied Math*, **4**(1), 73-81.

Giveen, S. M. (1963). A taxicab problem with time-dependent arrival rates. *SIAM Review*, **5**(2), 119-127.

Goh, C. H., Greenberg, B. S., Matsuo, H. (1993). Perishable inventory systems with batch demand and arrivals. *Operations Research Letters*, **13**(1), 1-8.

Graves, S. C. (1982). The application of queueing theory to continuous perishable inventory systems. *Management Science*, **28**(4), 400-406.

Gopikrishnan, P., Plerou, V., Gabaix, X., Stanley, H. E. (2000). Statistical properties of share volume traded in financial markets. *Physical Review E*, **62**(4), 44934496.

Guo, P., Lian, Z., Wang, Y. (2011). Pricing perishable products with compound Poisson demands. *Probability in the Engineering and Informational Sciences*, **25**(3), 289-306.

Hendershott, T., Mendelson, H. (2002). Crossing networks and dealer markets: competition and performance. *The Journal of Finance*, **55**(5), 2071-2115.

ITG SEC Filing - Required report on routing of customer orders for quarter ending June 2011. Retrieved from: *http://www.itg.com/order_routing/Rule606Report_Q2_2011.pdf*

Kashyap, B. R. K. (1966). The double-ended queue with bulk service and limited waiting space. *Operations Research*, **14**(5), 822-834.

Kaspi, H., Perry, D. (1983). Inventory systems of perishable commodities. *Advances in Applied Probability*, **15**, 674-685.

Kaspi H., Perry D. (1984). Inventory systems for perishable commodities with renewal input and Poisson output. *Advances in Applied Probability*, **16**, 402-421.

Kendall, D. G. (1951). Some problems in the theory of queues. *Journal of the Royal Statistical Society*, **B**(13), 151-185.

Kim, W. K., Yoon, K. P., Mendoza, G., Sedaghat, M. (2010). Simulation model for extended double-ended queueing. *Computers & Industrial Engineering*, **59**(2), 209-219.

Lee, J. (2007). First exit times for compound Poisson dams with a general release rule. *Mathematical Methods of Operations Research*, **65**(1), 169-178.

Maglaras, C., Moallemi, C. (2011). A multiclass model of limit order book dynamics and its application to optimal trade execution. Working paper, Columbia Business School.

Maslov, S., Mills, M. (2001). Price fluctuations from the order book perspective: empirical facts and a simple model. *Physica A: Statistical Mechanics and its Applications*, **299**(1), 234-246.

Mittal, H. (2008). Are you playing in a toxic dark pool? A guide to preventing information leakage, *Journal of Trading*, **3**(3), 20-33.

Nahmias, S., Perry, D., Stadje, W. (2004). Perishable inventory systems with variable input and demand rates. *Mathematical Methods of Operations Research*, **60**, 155-162.

Parlour, C. A. (1998). Price dynamics in limit order markets. *Review of Financial Studies*, **11**(4), 789-816.

Perry, D. (1985). An inventory system for perishable commodities with random lifetime. *Advances in Applied Probability*, **17**, 234-236.

Perry, D., Stadje, W. (1999). Perishable inventory systems with impatient demands. *Mathematical Methods of Operations Research*, **50**(1), 77-90.

Perry, D., Stadje, W., Zacks, S. (1999). First-exit times for increasing compound processes. *Stochastic Models*, **15**(5), 977-992.

Perry, D., Stadje, W., Zacks, S. (2002). First-exit times for compound Poisson processes for some types of positive and negative jumps. *Stochastic Models*, **18**(1), 139-157.

Ray, S. (2010). A match in the dark: Understanding crossing network liquidity. Working Paper. University of Florida, Warrington College of Business Administration.

Ready, M.J. (2009). Determinants of volume in dark pools. *AFA 2010 Atlanta Meetings Paper.*

Rosu, I. (2009). A dynamic model of the limit order book. *Review of Financial Studies*, **22**(11), 4601-4641.

Schack, J., Gawronski, J. (2011). Let there be light: Rosenblatt's monthly dark liquidity tracker *Rosenblatt Securities Inc.* February 28, 2011.

Sofianos, G. (2007). Dark pools and algorithmic trading. *Algorithmic Trading Handbook, 2nd Edition.*

Som, P., Wilhelm, W. E., Disney, R. L. (1994). Kitting process in a stochastic assembly system. *Queueing Systems*, **17**(3), 471-490.

Takahashi, M., Osawa, H., Fujisawa, T. (2000). On a synchronization queue with two finite buffers. *Queueing Systems*, **36**(1), 107-123.

Zacks, S. (2004). Generalized integrated telegraph processes and the distribution of related stopping times. *Journal of Applied Probability*, **41**(2), 497-507.

Zenios, S. A. (1999). Modeling the transplant waiting list: A queueing model with reneging. *Queueing Systems*, **31**(3), 239-251.

Zhu, H. (2012). Do dark pools harm price discovery? Working Paper. Massachusetts Institute of Technology, Sloan School of Management.

Appendix

Proof of Proposition 1: Level-crossing theory states that at any given y, the rate of up-crossing must equal the rate of down-crossing. This implies the following integral equation for $f(y)$:

$$
\begin{aligned}
&r(y)f(y) \\
&+\lambda_a\left[\int_y^\infty \exp\left\{-\mu_a(z-y)\right\} f(z)dz + \mathbb{P}_0 \exp\left\{\mu_a y\right\} 1_{\{y\leq 0\}}\right] + \left[\lambda_a^I \int_y^\infty \exp\left\{-\mu_a^I(z-y)\right\} f(z)dz\right] 1_{\{y>0\}} \\
&=\lambda_b\left[\int_{-\infty}^y \exp\left\{\mu_b(z-y)\right\} f(z)dz + \mathbb{P}_0 \exp\left\{-\mu_b y\right\} 1_{\{y\geq 0\}}\right] + \left[\lambda_b^I \int_{-\infty}^y \exp\left\{\mu_b^I(z-y)\right\} f(z)dz\right] 1_{\{y<0\}}.
\end{aligned}
$$

Or more compactly, taking the integral over the atom at 0,

$$
\begin{aligned}
r(y)f(y) + \lambda_a \int_y^\infty \exp\left\{-\mu_a(z-y)\right\} dF(z) + \left[\lambda_a^I \int_y^\infty \exp\left\{-\mu_a^I(z-y)\right\} dF(z)\right] 1_{\{y>0\}} \\
= \lambda_b \int_{-\infty}^y \exp\left\{\mu_b(z-y)\right\} dF(z) + \left[\lambda_b^I \int_{-\infty}^y \exp\left\{\mu_b^I(z-y)\right\} dF(z)\right] 1_{\{y<0\}}. \qquad (18)
\end{aligned}
$$

We derive (1) and (2) by successively taking derivatives of (18), transforming the integro-differential equation into a third-order differential equation. We do so for the case of $y>0$. The analogous result holds for $y<0$. Simplifying (18) for $y>0$,

$$
\begin{aligned}
r(y)f(y) + \lambda_a \int_y^\infty \exp\left\{-\mu_a(z-y)\right\} dF(z) + \lambda_a^I \int_y^\infty \exp\left\{-\mu_a^I(z-y)\right\} dF(z) \\
= \lambda_b \int_{-\infty}^y \exp\left\{\mu_b(z-y)\right\} dF(z). \qquad (19)
\end{aligned}
$$

Taking derivatives on both sides yields

$$
\begin{aligned}
f'(y)r(y) + f(y)r'(y) + \lambda_a\left(\mu_a \int_y^\infty \exp\left\{-\mu_a(z-y)\right\} dF(z) - f(y)\right) \\
+\lambda_a^I\left(\mu_a^I \int_y^\infty \exp\left\{-\mu_a^I(z-y)\right\} dF(z) - f(y)\right) \\
= \lambda_b\left(f(y) - \mu_b \int_{-\infty}^y \exp\left\{\mu_b(z-y)\right\} dF(z)\right).
\end{aligned}
$$

Rewriting this equation gives

$$
\begin{aligned}
\lambda_a \int_y^\infty \exp\left\{-\mu_a(z-y)\right\} dF(z) = \frac{1}{\mu_a}\Big[\lambda_b(f(y) - \mu_b \int_{-\infty}^y \exp\left\{\mu_b(z-y)\right\} dF(z)) \\
+ \lambda_a^I\Big(f(y) - \mu_a^I \int_y^\infty \exp\left\{-\mu_a^I(z-y)\right\} dF(z)\Big) + \lambda_a f(y) - f(y)r'(y) - f'(y)r(y)\Big].
\end{aligned}
$$

Substituting the LHS into (19), we get

$$
\begin{aligned}
f(y)\Big(\lambda_a + \mu_a r(y) - r'(y)\Big) - f'(y)r(y) = -f(y)(\lambda_b + \lambda_a^I) \\
+ \lambda_b(\mu_b + \mu_a) \int_{-\infty}^y \exp\left\{\mu_b(z-y)\right\} dF(z) + \lambda_a^I(\mu_a^I - \mu_a) \int_y^\infty \exp\left\{-\mu_a^I(z-y)\right\} dF(z). \qquad (20)
\end{aligned}
$$

Taking derivatives on both sides yields

$$-f''(y)r(y)-f'(y)r'(y)+f'(y)\Big(\lambda_a+\mu_a r(y)-r'(y)\Big)+f(y)\Big(\mu_a r'(y)-r''(y)\Big)=-f'(y)(\lambda_b+\lambda_a^I)$$
$$+\lambda_b(\mu_b+\mu_a)\Big(f(y)-\mu_b\int_{-\infty}^{y}\exp\{\mu_b(z-y)\}\,dF(z)\Big)+\lambda_a^I(\mu_a^I-\mu_a)\Big(\mu_a^I\int_{y}^{\infty}\exp\{-\mu_a^I(z-y)\}\,dF(z)-f(y)\Big).$$

Simplifying gives

$$\lambda_b(\mu_b+\mu_a)\int_{-\infty}^{y}\exp\{\mu_b(z-y)\}\,dF(z)=$$
$$\frac{1}{\mu_b}\Big[\lambda_a^I(\mu_a^I-\mu_a)\mu_a^I\int_{y}^{\infty}\exp\{-\mu_a^I(z-y)\}\,dF(z)+f''(y)r(y)+f'(y)\Big(2r'(y)-\mu_a r(y)-(\lambda_b+\lambda_a+\lambda_a^I)\Big)$$
$$+f(y)\Big(r''(y)-\mu_a r'(y)+\lambda_b(\mu_b+\mu_a)+\lambda_a^I(\mu_a-\mu_a^I)\Big)\Big].$$

Substituting the LHS into (20) yields

$$\begin{aligned}f(y)\Big(\lambda_a+\mu_a r(y)-r'(y)\Big)-r(y)f'(y)=&\frac{1}{\mu_b}\Big[\lambda_a^I(\mu_a^I-\mu_a)\mu_a^I\int_{y}^{\infty}\exp\{-\mu_a^I(z-y)\}\,dF(z)\\&+f''(y)r(y)+f'(y)\Big(2r'(y)-\mu_a r(y)-(\lambda_b+\lambda_a+\lambda_a^I)\Big)\\&+f(y)\Big(r''(y)-\mu_a r'(y)+\lambda_b(\mu_b+\mu_a)+\lambda_a^I(\mu_a-\mu_a^I)\Big)\Big]\\&+\lambda_a^I(\mu_a^I-\mu_a)\int_{y}^{\infty}\exp\{-\mu_a^I(z-y)\}\,dF(z)-f(y)(\lambda_b+\lambda_a^I).\end{aligned}$$

Simplifying gives

$$\begin{aligned}\lambda_a^I(\mu_a-\mu_a^I)(\mu_a^I+\mu_b)&\int_{y}^{\infty}\exp\{-\mu_a^I(z-y)\}\,dF(z)\\&=f''(y)r(y)+f'(y)\Big(2r'(y)+(\mu_b-\mu_a)r(y)-(\lambda_b+\lambda_a+\lambda_a^I)\Big)\\&\quad+f(y)\Big(r''(y)+(\mu_b-\mu_a)r'(y)-\mu_a\mu_b r(y)+\lambda_b\mu_a-\lambda_a\mu_b+\lambda_a^I(\mu_a-\mu_b-\mu_a^I)\Big).\end{aligned}\tag{21}$$

Taking derivatives we get

$$\begin{aligned}\lambda_a^I(\mu_a-\mu_a^I)(\mu_a^I+\mu_b)\Big(\mu_a^I\int_{y}^{\infty}&\exp\{-\mu_a^I(z-y)\}\,dF(z)-f(y)\Big)=\\&f'''(y)r(y)+f''(y)r'(y)+f''(y)\Big(2r'(y)+(\mu_b-\mu_a)r(y)-(\lambda_b+\lambda_a+\lambda_a^I)\Big)\\&\quad+f'(y)\Big(2r''(y)+(\mu_b-\mu_a)r'(y)-\mu_a\mu_b r(y)\Big)+f'(y)\Big(\lambda_b\mu_a-\lambda_a\mu_b+\lambda_a^I(\mu_a-\mu_b-\mu_a^I)\Big)\\&\quad+f'(y)\Big(2r''(y)+(\mu_b-\mu_a)r'(y)\Big)+f(y)\Big(r'''(y)+(\mu_b-\mu_a)r''(y)-\mu_a\mu_b r'(y)\Big).\end{aligned}$$

Simplifying gives

$$\begin{aligned}\lambda_a^I(\mu_a-\mu_a^I)(\mu_a^I+\mu_b)&\int_{y}^{\infty}\exp\{-\mu_a^I(z-y)\}\,dF(z)\\&=\frac{1}{\mu_a^I}\Big[f'''(y)r(y)+f''(y)\Big(3r'(y)+(\mu_b-\mu_a)r(y)-(\lambda_b+\lambda_a+\lambda_a^I)\Big)\\&\quad+f'(y)\Big(3r''(y)+2(\mu_b-\mu_a)r'(y)-\mu_a\mu_b r(y)+\lambda_b\mu_a-\lambda_a\mu_b+\lambda_a^I(\mu_a-\mu_b-\mu_a^I)\Big)\\&\quad+f(y)\Big(r'''(y)+(\mu_b-\mu_a)r''(y)-\mu_a\mu_b r'(y)+\lambda_a^I(\mu_a-\mu_a^I)(\mu_a^I+\mu_b)\Big].\end{aligned}$$

Substituting into (21) yields

$$\begin{aligned}
&\frac{1}{\mu_a^I}\Big[f'''(y)r(y)+f''(y)\Big(3r'(y)+(\mu_b-\mu_a)r(y)-(\lambda_b+\lambda_a+\lambda_a^I)\Big)\\
&\quad+f'(y)\Big(3r''(y)+2(\mu_b-\mu_a)r'(y)-\mu_a\mu_b r(y)+\lambda_b\mu_a-\lambda_a\mu_b+\lambda_a^I(\mu_a-\mu_b-\mu_a^I)\Big)\\
&\quad+f(y)\Big(r'''(y)+(\mu_b-\mu_a)r''(y)-\mu_a\mu_b r'(y)+\lambda_a^I(\mu_a-\mu_a^I)(\mu_a^I+\mu_b)\Big)\Big]\\
=&f''(y)r(y)+f'(y)\Big(2r'(y)+(\mu_b-\mu_a)r(y)-(\lambda_b+\lambda_a+\lambda_a^I)\Big)\\
&\quad+f(y)\Big(r''(y)+(\mu_b-\mu_a)r'(y)-\mu_a\mu_b r(y)+\lambda_b\mu_a-\lambda_a\mu_b+\lambda_a^I(\mu_a-\mu_b-\mu_a^I)\Big).
\end{aligned}$$

Collecting terms and simplifying, we get (1). Using a similar procedure, we can derive (2). □

Proof of Theorem 1: Note that all parameter values are positive.

For $i=\{a,b\}$, where for $j\in\{a,b\}$ and $j\neq i$, let

$$\begin{aligned}
A_i&=\frac{k_i}{\mu_i\mu_j\mu_j^I},\\
B_i&=k_i\left(\frac{1}{\mu_j\mu_j^I}-\frac{1}{\mu_i\mu_j^I}-\frac{1}{\mu_j\mu_i}\right)-\frac{\lambda_i+\lambda_j+\lambda_j^I}{\mu_i\mu_j\mu_j^I},\\
C_i&=k_i\left(\frac{1}{\mu_i}-\frac{1}{\mu_j}-\frac{1}{\mu_j^I}\right)+\lambda_i\left(\frac{\mu_j+\mu_j^I}{\mu_i\mu_j\mu_j^I}\right)+\lambda_j\left(\frac{\mu_j^I-\mu_i}{\mu_i\mu_j\mu_j^I}\right)+\lambda_j^I\left(\frac{\mu_j-\mu_i}{\mu_i\mu_j\mu_j^I}\right),\\
D_i&=k_i+\frac{\lambda_j}{\mu_j}-\frac{\lambda_i}{\mu_i}+\frac{\lambda_j^I}{\mu_j^I}.
\end{aligned}$$

Let

$$Z^l\triangleq\frac{d^l f(y)}{(dy)^l}$$

be the l^{th} derivative of $f(y)$. Then for $r(y)=k_b 1_{\{y>0\}}-k_a 1_{\{y<0\}}$, (1) can be written as

$$A_bZ^3+B_bZ^2+C_bZ+D_b=0, \tag{22}$$

and (2) can be written as

$$A_aZ^3-B_aZ^2+C_aZ-D_a=0. \tag{23}$$

Let Θ_{b_1}, Θ_{b_2}, and Θ_{b_3}, be the roots of (22) and Θ_{a_1}, Θ_{a_2}, and Θ_{a_3} be the roots of (23). Given the cubic form of (22), we know from the method of characteristic polynomials that $f(y)$ has the following form.

$$f(y)=\begin{cases}U_{a_1}e^{y\Theta_{a_1}}+U_{a_2}e^{y\Theta_{a_2}}+U_{a_3}e^{y\Theta_{a_3}} & y<0,\\ \mathbb{P}_0 & y=0,\\ U_{b_1}e^{y\Theta_{b_1}}+U_{b_2}e^{y\Theta_{b_2}}+U_{b_3}e^{y\Theta_{b_3}} & y>0.\end{cases}$$

The proof proceeds in several steps. In all steps we prove the results for the case for $y>0$. By symmetry all results hold for $y<0$.

- **Step 1.** We prove Lemma 1 that shows that when the stability condition (6) holds, all roots are real and that only one of them is negative (positive) for $y>0$ ($y<0$). We then argue that only the negative (positive) root can have a non-zero coefficient. This establishes that $f(y)$ has the form given in (7).
- **Step 2** We show how to determine the constants U_a, U_b, Θ_a and Θ_b.
- **Step 3** We establish that U_a,$U_b>0$. We do so by proving Lemma 2 which establishes a bound on the abandonment rates, k_a and k_b, the hold if (6) holds, and Lemma 3 that shows U_a and U_b are positive if this bound holds.
- **Step 4** We establish that $P_0>0$. We do so by considering a special case for which this relation holds and show that the general case is bounded below by it.

Step 1.

LEMMA 1. *(22) has one negative and two positive real roots iff the stability condition,* $k_b+\frac{\lambda_a^I}{\mu_a^I}>\frac{\lambda_b}{\mu_b}-\frac{\lambda_a}{\mu_a}$*, holds. Similarly (23) has two negative and one positive real roots iff* $k_a+\frac{\lambda_b^I}{\mu_b^I}>\frac{\lambda_a}{\mu_a}-\frac{\lambda_b}{\mu_b}$.

Proof of Lemma 1: We consider the case for $y>0$. Let $h(Z)=A_bZ^3+B_bZ^2+C_bZ+D_b$. Substituting in and simplifying we find

$$\begin{aligned} h(0) &= D_b \\ h\left(\mu_a^I\right) &= \frac{\lambda_a^I\left(\mu_a-\mu_a^I\right)\left(\mu_b+\mu_a^I\right)}{k_b}, \\ h\left(\mu_a\right) &= \frac{\lambda_a\left(\mu_a+\mu_b\right)\left(\mu_a^I-\mu_a\right)}{k_b}. \end{aligned}$$

We first consider the case if the stability condition holds. Observe by definition of D_b, if $k_b+\frac{\lambda_a^I}{\mu_a^I}>\frac{\lambda_b}{\mu_b}-\frac{\lambda_a}{\mu_a}$, then $h(0)>0$. If $\mu_a<\mu_a^I$, $h(\mu_a^I)<0$ and if $\mu_a>\mu_a^I$, $h(\mu_a)<0$. In either case, there must be a real root on the interval $(0,\max[\mu_a,\mu_a^I])$. If $\mu_a=\mu_a^I$, there is a positive real root at μ_a. Because $A_b>0$, $\lim_{Z\to\infty}h(Z)>0$. Therefore another positive real root exists in $(\max[\mu_a,\mu_a^I],\infty)$. Because $\lim_{Z\to-\infty}h(Z)<0$ one negative real root exists in$(-\infty,0)$. This proves the sufficiency of the stability condition. If there is one negative and two positive roots, then $\lim_{Z\to-\infty}h(Z)<0$ implies $h(0)=D_b>0$ implying the stability conditions holds. By symmetry the result holds for $y<0$. $\square$.

As a consequence of Lemma 1, in order for $f(y)$ to remain well-defined in the limit as $y\to\infty$, letting Θ_{b_2} and Θ_{b_3} be the positive roots, it must be $U_{b_2}=0$ and $U_{b_3}=0$.

Similarly, for $y<0$, if (6) holds, then we can show that there exists one positive and two negative real roots. Let Θ_{a_2} and Θ_{a_3} be the latter two. Then for $f(y)$ to remain well-define as $y\to-\infty$, it must be $U_{a_2}=U_{a_3}=0$.

Thus, the steady-state distribution of the queue length simplifies to

$$f(y) = \begin{cases} U_a\, e^{y\Theta_a} & y < 0 \\ \mathbb{P}_0 & y = 0 \\ U_b\, e^{-y\Theta_b} & y > 0 \end{cases} \tag{24}$$

where $\Theta_a \triangleq \Theta_{a_1} > 0$ and $\Theta_b \triangleq -\Theta_{b_1} > 0$ and $U_a \triangleq U_{a_1}$ and $U_b \triangleq U_{b_1}$.

Step 2. We note that Θ_a and Θ_b can be expressed in closed-form. Let

$$R_i = \frac{2B_i^3 - 9A_iB_iC_i + 27A_i^2D_b}{54A_i^3},\ P_i = \frac{B_i^2 - 3A_iC_i}{9A_i^2},\ \text{and}\ \theta_i = \text{Cos}^{-1}\left(\frac{R_i}{\sqrt{P_i^3}}\right)\ \text{for}\ i \in \{a, b\}.$$

Then, from e.g., Pachner (1983),

$$\begin{aligned} \Theta_a &= -2\sqrt{P_a}\text{Cos}\left(\frac{\theta_a + 2\pi}{3}\right) - \frac{B_a}{3A_a} \\ \Theta_b &= 2\sqrt{P_b}\,\text{Cos}\left(\frac{\theta_b}{3}\right) + \frac{B_b}{3A_b}. \end{aligned} \tag{25}$$

Using (24), the flow balance equations, (4) and (5), can be reduced to

$$\begin{aligned} \frac{\lambda_a}{\mu_a} &= \frac{\lambda_a U_b}{\Theta_b^2 + \Theta_b\mu_a} + \frac{\lambda_b U_a}{\Theta_a^2 + \Theta_a\mu_b} + \frac{\lambda_b^I U_a}{\Theta_a^2 + \Theta_a\mu_b^I} + \frac{k_a U_a}{\Theta_a}, \\ \frac{\lambda_b}{\mu_b} &= \frac{\lambda_a U_b}{\Theta_b^2 + \Theta_b\mu_a} + \frac{\lambda_b U_a}{\Theta_a^2 + \Theta_a\mu_b} + \frac{\lambda_a^I U_b}{\Theta_b^2 + \Theta_b\mu_a^I} + \frac{k_b U_b}{\Theta_b}. \end{aligned}$$

Solving these two equations for U_a and U_b provide the constants given in the theorem. Then normalization condition, (3), can be reduced to

$$\mathbb{P}_0 = 1 - \frac{U_a}{\Theta_a} - \frac{U_b}{\Theta_b}.$$

Step 3. Next we show $U_a,\, U_b > 0$. Let

$$\kappa_a = \frac{\lambda_b^I}{\Theta_a + \mu_b^I}\ \text{and}\ \kappa_b = \frac{\lambda_a^I}{\Theta_b + \mu_a^I}.$$

We require the following lemma.

LEMMA 2. *If* $k_b + \frac{\lambda_a^I}{\mu_a^I} > \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$, *then* $k_b + \kappa_b > \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$. *Similarly, if* $k_a + \frac{\lambda_b^I}{\mu_b^I} > \frac{\lambda_a}{\mu_a} - \frac{\lambda_b}{\mu_b}$, *then* $k_a + \kappa_a > \frac{\lambda_a}{\mu_a} - \frac{\lambda_b}{\mu_b}$.

Proof of Lemma 2: We consider the first claim. If $k_b \geq \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$ then $k_b + \kappa_b > \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$ is immediate as $\kappa_b > 0$. Consider then the case $k_b < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$. Rewriting $k_b + \kappa_b > \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$ as $\frac{\lambda_a^I}{\frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a} - k_b} - \mu_a^I > \Theta_b$ and substituting Θ_b from (25), and simplifying, we need to show

$$\Pi > \text{Cos}\left(\frac{\theta_b}{3}\right). \tag{26}$$

where

$$\Pi = \frac{\frac{1}{2}\left(\frac{\lambda_a^I}{\frac{\lambda_b}{\mu_b}-\frac{\lambda_a}{\mu_a}-k_b} + \frac{\lambda_a+\lambda_b+\lambda_a^I+k_b\mu_a-k_b\mu_b-2k_b\mu_a^I}{3k_b}\right)}{\sqrt{P_b}}.$$

Using $k_b + \frac{\lambda_a^I}{\mu_a^I} > \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$ and $k_b < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$ we can show the numerator of the Π is positive. By Lemma 1, because (22) has three real roots, P_b is real and positive. Thus Π is positive.

If $\Pi > 1$, (26) holds trivially. Suppose then that $\Pi \leq 1$. We observe R_b is real and $P_b^3 > R_b^2$, holds. This implies that $0 < \theta_b < \pi$ which means $1/2 < \mathrm{Cos}(\theta_b/3) < 1$. Noting $\mathrm{Cos}(\theta/3)$ is decreasing in $0 < \theta < \pi$, taking Cos^{-1} of both sides of (26) gives $\mathrm{Cos}^{-1}(\Pi) < \theta_b/3$. Substituting for the value of θ_b and simplifying, we get $3\,\mathrm{Cos}^{-1}(\Pi) < \mathrm{Cos}^{-1}\left(R_b/\sqrt{P_b^3}\right)$. Noting that $\mathrm{Cos}\left(\mathrm{Cos}^{-1}(t)\right) = t$ for $|t| \leq 1$, we take the cosine of both sides, $\mathrm{Cos}\left(3\,\mathrm{Cos}^{-1}(\Pi)\right) > R_b/\sqrt{P_b^3}$. Utilizing the identity $\mathrm{Cos}(3t) = 4\mathrm{Cos}^3(t) - 3\mathrm{Cos}(t)$ gives $4\Pi^3 - 3\Pi > R_b/\sqrt{P_b^3}$. Substituting the definitions of Π, P_b and R_b and simplifying, results in

$$\frac{\lambda_a^I(\mu_a\mu_b)^3(\mu_a^I)^2\left(k_b+\frac{\lambda_a^I}{\mu_a^I}+\frac{\lambda_a}{\mu_a}-\frac{\lambda_b}{\mu_b}\right)\left(\left(\frac{\lambda_b}{\mu_b}-\frac{\lambda_a}{\mu_a}\right)\left(k_b+\frac{\lambda_a^I}{\mu_a^I}+\frac{\lambda_a}{\mu_a}-\frac{\lambda_b}{\mu_b}\right)+\frac{\mu_a\mu_b}{\mu_a^I}\left(\frac{\lambda_b}{\mu_b}-\frac{\lambda_a}{\mu_a}-k_b\right)\left(\frac{\lambda_b}{\mu_b^2}+\frac{\lambda_a}{\mu_a^2}\right)\right)}{2k_b\mu_a^3\mu_b^3\left(\frac{\lambda_b}{\mu_b}-\frac{\lambda_a}{\mu_a}-k_b\right)^3} > 0.$$

The assumption that $k_b < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$ and the stability condition $k_b + \frac{\lambda_a^I}{\mu_a^I} > \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$ establish that the left hand side of the above equation is indeed positive and (26) holds. By symmetry the second claimed result holds. □

LEMMA 3. *If* $-k_a - \frac{\lambda_b^I}{\mu_b^I} < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a} < \frac{\lambda_a^I}{\mu_a^I} + k_b$, *then* $U_a, U_b > 0$ *and* $\mathbb{P}_0 < 1$.

Proof of Lemma 3: From its definition, $U_a > 0$ if $k_b\mu_b\Theta_b^2 + \left(\left(\lambda_a + \lambda_a^I + k_b\left(\mu_a + \mu_a^I\right)\right)\mu_b - \lambda_b\mu_a\right)\Theta_b - \lambda_b\mu_a\mu_a^I + \left(\lambda_a^I\mu_a + \left(\lambda_a + k_b\mu_a\right)\mu_a^I\right)\mu_b > 0$. Observe

$$\begin{aligned}
&k_b\mu_b\Theta_b^2 + \left(\left(\lambda_a+\lambda_a^I+k_b\left(\mu_a+\mu_a^I\right)\right)\mu_b - \lambda_b\mu_a\right)\Theta_b - \lambda_b\mu_a\mu_a^I + \left(\lambda_a^I\mu_a + \left(\lambda_a+k_b\mu_a\right)\mu_a^I\right)\mu_b \\
&= \lambda_a^I\mu_b\mu_a\left(k_b\left(\frac{\Theta_b+\mu_a}{\mu_a}\right)\left(\frac{\Theta_b+\mu_a^I}{\lambda_a^I}\right) + \frac{\lambda_a}{\mu_a}\left(\frac{\Theta_b+\mu_a^I}{\lambda_a^I}\right) + \frac{\Theta_b+\mu_a}{\mu_a} - \frac{\lambda_b}{\mu_b}\left(\frac{\Theta_b+\mu_a^I}{\lambda_a^I}\right)\right) \\
&= \frac{\lambda_a^I\mu_b\mu_a}{\kappa_b}\left((k_b+\kappa_b)\left(1+\frac{\Theta_b}{\mu_a}\right) - \frac{\lambda_b}{\mu_b} + \frac{\lambda_a}{\mu_a}\right) \\
&> \frac{\lambda_a^I\mu_b\mu_a}{\kappa_b}\left(k_b+\kappa_b-\frac{\lambda_b}{\mu_b}+\frac{\lambda_a}{\mu_a}\right) \\
&> 0
\end{aligned}$$

where the last inequality follows from Lemma 2. Similarly, one can show $U_b > 0$. As a result, $\mathbb{P}_0 < 1$ since $\mathbb{P}_0 = 1 - \frac{U_a}{\Theta_a} - \frac{U_b}{\Theta_b}$. □

Step 4. It remains to show that if $-k_a - \frac{\lambda_b^I}{\mu_b^I} < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a} < \frac{\lambda_a^I}{\mu_a^I} + k_b$ then $\mathbb{P}_0 > 0$. We proceed as follows: We first consider an instance of the queue where there are no impatient customers. Designating this the "tilde" system, we show in Lemma 4 that if the system is stable then $\tilde{\mathbb{P}}_0 > 0$

We then define a system where there are only impatient customers, designated the "hat" system. We show in Lemma 5 by appropriate choice of parameters that the two systems can be made equivalent in performance. Using these lemmas we establish in Lemma 6 that if the original system is stable, there is a positive lower bound on P_0.

Consider the case with no impatient customers ($\lambda_a^I = \lambda_b^I = 0$, $\omega_a^I = \omega_b^I = 0$) and abandonment rates $\tilde{k}_a$ and $\tilde{k}_b$. In this "tilde" system, the (1) and (2) reduce to

$$\begin{aligned}
&\tilde{f}''(y) + \tilde{f}'(y)\left(\frac{\tilde{k}_b(\mu_b-\mu_a)-(\lambda_a+\lambda_b)}{\tilde{k}_b}\right) + \tilde{f}(y)\left(\frac{\lambda_b\mu_a-\lambda_a\mu_b-\mu_b\mu_a\tilde{k}_b}{\tilde{k}_b}\right) = 0 \qquad \text{if } y>0,\\
&\tilde{f}''(y) + \tilde{f}'(y)\left(\frac{\tilde{k}_a(\mu_b-\mu_a)+(\lambda_a+\lambda_b)}{\tilde{k}_a}\right) + f(y)\left(\frac{\lambda_a\mu_b-\lambda_b\mu_a-\mu_b\mu_a\tilde{k}_a}{\tilde{k}_a}\right) = 0 \qquad \text{if } y<0.
\end{aligned} \tag{27}$$

The solution to (27) (i.e., $\tilde{f}(y)$) is as in (7)

$$\tilde{f}(y) = \begin{cases} \tilde{U}_a\, e^{y\tilde{\Theta}_a} & y<0,\\ \tilde{\mathbb{P}}_0 & y=0,\\ \tilde{U}_b\, e^{-y\tilde{\Theta}_b} & y>0, \end{cases}$$

where

$$\begin{aligned}
\tilde{\Theta}_a &= \tilde{\beta}_a - \tilde{\alpha}_a,\ \tilde{\Theta}_b = \tilde{\beta}_b - \tilde{\alpha}_b\\
\tilde{\alpha}_a &= \frac{\lambda_a+\lambda_b}{2\tilde{k}_a} + \frac{\mu_b-\mu_a}{2},\ \tilde{\beta}_a = \sqrt{\tilde{\alpha}_a^2 + \frac{\lambda_b\mu_a-\lambda_a\mu_b}{\tilde{k}_a} + \mu_a\mu_b},\\
\tilde{\alpha}_b &= \frac{\lambda_a+\lambda_b}{2\tilde{k}_b} + \frac{\mu_a-\mu_b}{2},\ \tilde{\beta}_b = \sqrt{\tilde{\alpha}_b^2 + \frac{\lambda_a\mu_b-\lambda_b\mu_a}{\tilde{k}_b} + \mu_a\mu_b},
\end{aligned}$$

and

$$\begin{aligned}
\tilde{U}_a &= \frac{\lambda_a\tilde{\Theta}_a(\tilde{\Theta}_a+\mu_b)(\lambda_a\mu_b-\lambda_b\mu_a+\tilde{k}_b\mu_b(\tilde{\Theta}_b+\mu_a))}{\mu_a\mu_b(\tilde{k}_b\lambda_b(\tilde{\Theta}_b+\mu_a)+\tilde{k}_a(\lambda_a+\tilde{k}_b(\tilde{\Theta}_b+\mu_a))(\tilde{\Theta}_a+\mu_b))},\\
\tilde{U}_b &= \frac{\lambda_b\tilde{\Theta}_b(\tilde{\Theta}_b+\mu_a)(\lambda_b\mu_a-\lambda_a\mu_b+\tilde{k}_a\mu_a(\tilde{\Theta}_a+\mu_b))}{\mu_a\mu_b(\tilde{k}_b\lambda_b(\tilde{\Theta}_b+\mu_a)+\tilde{k}_a(\lambda_a+\tilde{k}_b(\tilde{\Theta}_b+\mu_a))(\tilde{\Theta}_a+\mu_b))},\\
\tilde{\mathbb{P}}_0 &= 1-\frac{\tilde{U}_a}{\tilde{\Theta}_a}-\frac{\tilde{U}_b}{\tilde{\Theta}_b}.
\end{aligned}$$

We show

LEMMA 4. *If* $-\tilde{k}_a < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a} < \tilde{k}_b$, $\tilde{\mathbb{P}}_0 > 0$.

Proof of Lemma 4: Substituting in $\tilde{\Theta}_a$, $\tilde{\Theta}_b$, $\tilde{U}_a$, $\tilde{U}_b$ into the definition of $\tilde{\mathbb{P}}_0$ and reducing implies $\tilde{\mathbb{P}}_0 > 0$ if

$$\frac{\mu_a\mu_b\left(\tilde{\Theta}_b+\mu_a\right)\left(\tilde{\Theta}_a+\mu_b\right)\left(\tilde{k}_a-\left(\frac{\lambda_a}{\mu_a}-\frac{\lambda_b}{\mu_b}\right)\right)\left(\tilde{k}_b-\left(\frac{\lambda_b}{\mu_b}-\frac{\lambda_a}{\mu_a}\right)\right)}{\mu_a\mu_b(\tilde{k}_b\lambda_b(\tilde{\Theta}_b+\mu_a)+\tilde{k}_a(\lambda_a+\tilde{k}_b(\tilde{\Theta}_b+\mu_a))(\tilde{\Theta}_a+\mu_b))}\left(1-\frac{\frac{\lambda_b}{\mu_b}\left(\frac{\tilde{\Theta}_a}{\tilde{\Theta}_a+\mu_b}\right)}{\tilde{k}_a-\left(\frac{\lambda_a}{\mu_a}-\frac{\lambda_b}{\mu_b}\right)}-\frac{\frac{\lambda_a}{\mu_a}\left(\frac{\tilde{\Theta}_b}{\tilde{\Theta}_b+\mu_a}\right)}{\tilde{k}_b-\left(\frac{\lambda_b}{\mu_b}-\frac{\lambda_a}{\mu_a}\right)}\right)>0$$

Then $-\tilde{k}_a < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a} < \tilde{k}_b$, implies that the above equation holds when

$$1 - \frac{\frac{\lambda_b}{\mu_b}\left(\frac{\tilde{\Theta}_a}{\tilde{\Theta}_a+\mu_b}\right)}{\tilde{k}_a - \left(\frac{\lambda_a}{\mu_a} - \frac{\lambda_b}{\mu_b}\right)} - \frac{\frac{\lambda_a}{\mu_a}\left(\frac{\tilde{\Theta}_b}{\tilde{\Theta}_b+\mu_a}\right)}{\tilde{k}_b - \left(\frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}\right)} = 1 - \frac{h(p)}{p} - \frac{g(q)}{q} > 0$$

where $p = \tilde{k}_a - (\frac{\lambda_a}{\mu_a} - \frac{\lambda_b}{\mu_b})$, $q = \tilde{k}_b - (\frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a})$ and

$$h(p) = \frac{p}{2} + \frac{\frac{\lambda_b\mu_a}{\mu_b} + \frac{\lambda_a\mu_b}{\mu_a}}{2(\mu_a+\mu_b)} - \frac{1}{2p(\mu_a+\mu_b)}\sqrt{\frac{\lambda_b^2}{\mu_b^2}\mu_a^2 + \mu_b^2\left(p + p\frac{\mu_a}{\mu_b} + \frac{\lambda_a}{\mu_a}\right)^2 - 2\lambda_b\mu_a\left(p + p\frac{\mu_a}{\mu_b} - \frac{\lambda_a}{\mu_a}\right)},$$

$$g(q) = \frac{q}{2} + \frac{\frac{\lambda_b\mu_a}{\mu_b} + \frac{\lambda_a\mu_b}{\mu_a}}{2(\mu_a+\mu_b)} - \frac{1}{2q(\mu_a+\mu_b)}\sqrt{\frac{\lambda_b^2}{\mu_b^2}\mu_a^2 + \mu_b^2\left(q + q\frac{\mu_a}{\mu_b} - \frac{\lambda_a}{\mu_a}\right)^2 + 2\lambda_b\mu_a\left(q + q\frac{\mu_a}{\mu_b} + \frac{\lambda_a}{\mu_a}\right)}.$$

After some algebra, $\frac{h(p)}{p} + \frac{g(q)}{q} < 1$ holds if

$$p\sqrt{\frac{\lambda_b^2}{\mu_b^2}\mu_a^2 + \mu_b^2\left(p + p\frac{\mu_a}{\mu_b} + \frac{\lambda_a}{\mu_a}\right)^2 - 2\lambda_b\mu_a\left(p + p\frac{\mu_a}{\mu_b} - \frac{\lambda_a}{\mu_a}\right)}$$
$$+q\sqrt{\frac{\lambda_b^2}{\mu_b^2}\mu_a^2 + \mu_b^2\left(q + q\frac{\mu_a}{\mu_b} - \frac{\lambda_a}{\mu_a}\right)^2 + 2\lambda_b\mu_a\left(q + q\frac{\mu_a}{\mu_b} + \frac{\lambda_a}{\mu_a}\right)} > (p+q)\left(\frac{\frac{\lambda_b\mu_a}{\mu_b} + \frac{\lambda_a\mu_b}{\mu_a}}{2(\mu_a+\mu_b)}\right).$$

Observing both sides are positive, we can raise both sides to the power of 4. Subtracting and simplifying implies that the result holds if $4(p+q)^2\lambda_a\lambda_b(\mu_a+\mu_b)^2 > 0$, which is evident. □

Now consider the system with $k_a = k_b = 0$ (no patient customers) and impatient customers with parameters $\hat{\lambda}_a^I$, $\hat{\lambda}_b^I$, $\hat{\mu}_a^I$, and $\hat{\mu}_b^I$. Designating this the "hat" system, (1) and (2) reduce to two homogeneous, constant-coefficient second-order differential equations

$$\begin{aligned}
&\hat{f}''(y) + \hat{f}'(y)\left(\frac{\hat{\lambda}_a^I(\mu_b-\mu_a) + \lambda_a(\mu_b-\hat{\mu}_a^I) - \lambda_b(\mu_a+\hat{\mu}_a^I)}{\lambda_a+\lambda_b+\hat{\lambda}_a^I}\right) \\
&\qquad + \hat{f}(y)\left(\frac{-\hat{\lambda}_a^I\mu_a\mu_b - \lambda_a\hat{\mu}_a^I\mu_b + \lambda_b\mu_a\hat{\mu}_a^I}{\lambda_a+\lambda_b+\hat{\lambda}_a^I}\right) = 0 \qquad \text{if } y > 0, \\
&\hat{f}''(y) + \hat{f}'(y)\left(\frac{\hat{\lambda}_b^I(\mu_b-\mu_a) + \lambda_b(\hat{\mu}_b^I-\mu_a) + \lambda_a(\mu_b+\hat{\mu}_b^I)}{\lambda_a+\lambda_b+\hat{\lambda}_b^I}\right) \\
&\qquad + \hat{f}(y)\left(\frac{\lambda_a\mu_b\hat{\mu}_b^I - \hat{\lambda}_b^I\mu_a\mu_b - \lambda_b\mu_a\hat{\mu}_b^I}{\lambda_a+\lambda_b+\hat{\lambda}_b^I}\right) = 0 \qquad \text{if } y < 0.
\end{aligned} \tag{28}$$

The solution to (28), i.e., $\hat{f}(y)$, is as in (7) where:

$$\hat{\Theta}_a = \hat{\beta}_a - \hat{\alpha}_a, \;\; \hat{\Theta}_b = \hat{\beta}_b - \hat{\alpha}_b$$

$$\hat{\alpha}_a = \frac{\lambda_b^I(\mu_b-\mu_a) + \lambda_b(\mu_b^I-\mu_a) + \lambda_a(\mu_b+\mu_b^I)}{2(\lambda_a+\lambda_b+\lambda_b^I)}, \;\; \hat{\beta}_a = \sqrt{\hat{\alpha}_a^2 + \frac{\mu_a\mu_b\mu_b^I}{\lambda_a+\lambda_b+\lambda_b^I}\left(\frac{\lambda_b}{\mu_b} + \frac{\lambda_b^I}{\mu_b^I} - \frac{\lambda_a}{\mu_a}\right)},$$

$$\hat{\alpha}_b = \frac{\lambda_a^I(\mu_a-\mu_b) + \lambda_a(\mu_a^I-\mu_b) + \lambda_b(\mu_a+\mu_a^I)}{2(\lambda_a+\lambda_a^I+\lambda_b)}, \;\; \hat{\beta}_b = \sqrt{\hat{\alpha}_b^2 + \frac{\mu_a\mu_b\mu_a^I}{\lambda_a^I+\lambda_a+\lambda_b}\left(\frac{\lambda_a}{\mu_a} + \frac{\lambda_a^I}{\mu_a^I} - \frac{\lambda_b}{\mu_b}\right)},$$

and

$$\hat{U}_a = \frac{\frac{\lambda_a}{\mu_a}\hat{\Theta}_a\left(\hat{\Theta}_a+\mu_b\right)\left(\hat{\Theta}_a+\mu_b^I\right)\left(\lambda_a^I\mu_a+\lambda_a\mu_a^I-\frac{\lambda_b}{\mu_b}\mu_a\mu_a^I+\hat{\Theta}_b\left(\lambda_a+\lambda_a^I-\frac{\lambda_b}{\mu_b}\mu_a\right)\right)}{\hat{\Theta}_a\hat{\Theta}_b\left(\lambda_a\lambda_b^I+\lambda_a^I\left(\lambda_b+\lambda_b^I\right)\right)+\hat{\Theta}_a\left(\lambda_a^I\left(\lambda_b+\lambda_b^I\right)\mu_a+\lambda_a\lambda_b^I\mu_a^I\right)+\hat{\Theta}_b\left(\left(\lambda_a+\lambda_a^I\right)\lambda_b^I\mu_b+\lambda_a^I\lambda_b\mu_b^I\right)+\lambda_a\lambda_b^I\mu_a^I\mu_b+\lambda_a^I\mu_a\left(\lambda_b^I\mu_b+\lambda_b\mu_b^I\right)},$$

$$\hat{U}_b = \frac{\frac{\lambda_b}{\mu_b}\hat{\Theta}_b\left(\hat{\Theta}_b+\mu_a\right)\left(\hat{\Theta}_b+\mu_a^I\right)\left(\lambda_b^I\mu_b+\lambda_b\mu_b^I-\frac{\lambda_a}{\mu_a}\mu_b\mu_b^I+\hat{\Theta}_a\left(\lambda_b+\lambda_b^I-\frac{\lambda_a}{\mu_a}\mu_b\right)\right)}{\hat{\Theta}_a\hat{\Theta}_b\left(\lambda_a\lambda_b^I+\lambda_a^I\left(\lambda_b+\lambda_b^I\right)\right)+\hat{\Theta}_a\left(\lambda_a^I\left(\lambda_b+\lambda_b^I\right)\mu_a+\lambda_a\lambda_b^I\mu_a^I\right)+\hat{\Theta}_b\left(\left(\lambda_a+\lambda_a^I\right)\lambda_b^I\mu_b+\lambda_a^I\lambda_b\mu_b^I\right)+\lambda_a\lambda_b^I\mu_a^I\mu_b+\lambda_a^I\mu_a\left(\lambda_b^I\mu_b+\lambda_b\mu_b^I\right)},$$

$$\hat{\mathbb{P}}_0 = 1-\frac{\hat{U}_a}{\hat{\Theta}_a}-\frac{\hat{U}_b}{\hat{\Theta}_b}.$$

The stability condition now reduces to $-\frac{\lambda_b^I}{\mu_b^I} < \frac{\lambda_b}{\mu_b}-\frac{\lambda_a}{\mu_a} < \frac{\lambda_a^I}{\mu_a^I}$.

LEMMA 5. *If $\tilde{k}_i = \hat{\lambda}_j^I/(\hat{\Theta}_i+\hat{\mu}_j^I)$ for $i=\{a,b\}$, then $\hat{f}=\tilde{f}$ and $\hat{\mathbb{P}}_0=\tilde{\mathbb{P}}_0$.*

Proof of Lemma 5: For $y>0$, setting $\lambda_a^I=\lambda_b^I=\omega_a^I=\omega_b^I=0$ and re-arranging (18) implies the "tilde system" satisfies

$$\tilde{r}(y)\tilde{f}(y) = \lambda_b\int_{-\infty}^{y}\exp\left\{\mu_b(z-y)\right\}\tilde{f}(z)dz+\tilde{\mathbb{P}}_0\exp\left\{-\mu_b y\right\}-\lambda_a\int_{y}^{\infty}\exp\left\{-\mu_a(z-y)\right\}\tilde{f}(z)dz. \quad (29)$$

For $y>0$, setting $r(y)=0$ and re-arranging (18) implies the "hat system" satisfies

$$\begin{aligned}&\hat{\lambda}_a^I\int_{y}^{\infty}\exp\left\{-\hat{\mu}_a^I(z-y)\right\}\hat{f}(z)dz\\ &=\lambda_b\int_{-\infty}^{y}\exp\left\{\mu_b(z-y)\right\}\hat{f}(z)dz+\hat{\mathbb{P}}_0\exp\left\{-\mu_b y\right\}-\lambda_a\int_{y}^{\infty}\exp\left\{-\mu_a(z-y)\right\}\hat{f}(z)dz. \quad (30)\end{aligned}$$

Substituting $\hat{f}(z)=\hat{U}_b e^{-\hat{\Theta}_b z}$ into the left hand side of (30), and integrating implies for $y>0$,

$$\frac{\hat{\lambda}_a^I}{\hat{\Theta}_b+\hat{\mu}_a^I}\hat{U}_b e^{-\hat{\Theta}_b y} = \lambda_b\int_{-\infty}^{y}\exp\left\{\mu_b(z-y)\right\}\hat{f}(z)dz+\hat{\mathbb{P}}_0\exp\left\{-\mu_b y\right\}-\lambda_a\int_{y}^{\infty}\exp\left\{-\mu_a(z-y)\right\}\hat{f}(z)dz. \quad (31)$$

Observe that because the right hand sides of (29) and (31) have the same form, letting $\tilde{k}_b=\hat{\lambda}_a^I/(\hat{\Theta}_b+\hat{\mu}_a^I)$, then $\tilde{f}(y)=\hat{f}(y)$ solves (29). By symmetry the same argument holds for $y<0$. Therefore for $y>0$ and $y<0$, $\hat{f}=\tilde{f}$. From the normalization constraint, $\hat{\mathbb{P}}_0=\tilde{\mathbb{P}}_0$. □

Consider a system, designated "double-hat" in which $\hat{\hat{k}}_a=\hat{\hat{k}}_b=0$, $\hat{\hat{\mu}}_i^I \triangleq \mu_i^I$, and

$$\frac{\hat{\hat{\lambda}}_i^I}{\hat{\hat{\mu}}_i^I} \triangleq k_j+\frac{\lambda_i^I}{\mu_i^I}$$

for $i=\{a,b\}$, $j\neq i$. That is, there is no abandonment in the double-hat system and the ratio of the $\hat{\hat{\lambda}}_i^I$ to $\hat{\hat{\mu}}_i^I$ is fixed. Let $\hat{\hat{\mathbb{P}}}_0$ be the associated probability that this system is empty. By Lemma 5 there exists some equivalent system, say the "double-tilde" system with abandonment $\tilde{\tilde{k}}_i \triangleq \hat{\hat{\lambda}}_i^I/(\hat{\hat{\Theta}}_j+\hat{\hat{\mu}}_i^I)>0$ and no impatient customers (i.e., $\tilde{\tilde{\lambda}}_i^I=\tilde{\tilde{\omega}}_i^I>0$). (The definition of $\hat{\hat{\Theta}}_i$, $i\in\{a,b\}$ is given by $\hat{\Theta}_i$ above, substituting $\hat{\hat{\lambda}}_i^I$ for λ_i^I and $\hat{\hat{\mu}}_i^I$ for μ_i^I, $i\in\{a,b\}$.) We show first that $\hat{\hat{P}}_0>0$ and then establish it is a lower bound on P_0.

LEMMA 6. *If* $-k_a - \frac{\lambda_b^I}{\mu_b^I} < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a} < \frac{\lambda_a^I}{\mu_a^I} + k_b$, $\hat{\tilde{P}}_0 > 0$.

Proof of Lemma 6: By Lemma 5, $\hat{\tilde{P}}_0 = \tilde{\tilde{P}}_0$. We require

$$-\tilde{\tilde{k}}_a < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a} < \tilde{\tilde{k}}_b. \tag{32}$$

We show the second inequality in (32) holds. (By symmetry the first also will hold). If $\lambda_b/\mu_b - \lambda_a/\mu_a < 0$ then it holds trivially. Assume $\lambda_b/\mu_b - \lambda_a/\mu_a > 0$. Then we must show

$$\tilde{\tilde{k}}_b \triangleq \frac{\hat{\tilde{\lambda}}_a^I}{\hat{\tilde{\Theta}}_b + \hat{\tilde{\mu}}_a^I} = \frac{\lambda_a^I + k_b\mu_a^I}{\hat{\tilde{\Theta}}_b + \mu_a^I} > \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}$$

Rearranging, we must show

$$\frac{\mu_a^I\left(k_b + \frac{\lambda_a^I}{\mu_a^I} - \left(\frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}\right)\right)}{\frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}} > \hat{\tilde{\Theta}}_b$$

Substituting the definitions of $\hat{\tilde{\lambda}}_b^I$ and $\hat{\tilde{\mu}}_b^I$ into $\hat{\tilde{\Theta}}_b$, squaring both sides and rearranging gives

$$\frac{\mu_a^I(\lambda_a^I + k_b\mu_a^I)\left(k_b + \frac{\lambda_a^I}{\mu_a^I} - \left(\frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}\right)\right)\left(k_b + \frac{\lambda_a^I}{\mu_a^I} - \left(\frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}\right) + \frac{\lambda_b\mu_a}{\mu_b\mu_a^I} + \frac{\lambda_a\mu_b}{\mu_a\mu_a^I}\right)}{\mu_a\mu_b\left(\frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a}\right)^2(\lambda_a + \lambda_b + \lambda_a^I + k_b\mu_a^I)} > 0. \tag{33}$$

Under the lemma's assumption, (33) holds and therefore so does (32). By Lemma 4, $\tilde{\tilde{P}}_0 > 0$. □

Now consider the original system. For $y > 0$, the level-crossing equation for the general model is

$$\begin{aligned} r(y)f(y) + \lambda_a^I \int_y^\infty \exp\left\{-\mu_a^I(z-y)\right\} f(z)dz \\ = \lambda_b \int_{-\infty}^y \exp\left\{\mu_b(z-y)\right\} f(z)dz + \mathbb{P}_0 \exp\left\{-\mu_b y\right\} - \lambda_a \int_y^\infty \exp\left\{-\mu_a(z-y)\right\} f(z)dz. \end{aligned} \tag{34}$$

Since both (30) and (34) have the same form on the right-hand side, they will have equivalent behavior (i.e., $f(y) = \hat{f}(y)$ and $\mathbb{P}_0 = \hat{\mathbb{P}}_0$) if there exists $\hat{\lambda}_i^I$ and $\hat{\mu}_i^I$ such that

$$\hat{\lambda}_a^I \int_y^\infty \exp\left\{-\hat{\mu}_a^I(z-y)\right\} f(z)dz = r(y)f(y) + \lambda_a^I \int_y^\infty \exp\left\{-\mu_a^I(z-y)\right\} f(z)dz. \tag{35}$$

For $y > 0$, substituting $f(y) = U_b e^{-\Theta_b y}$ on both sides, we find (35) holds if

$$\frac{\hat{\lambda}_a^I}{\Theta_b + \hat{\mu}_a^I} = k_b + \frac{\hat{\lambda}_a}{\Theta_b + \mu_a^I}$$

Letting $\hat{\mu}_a^I \triangleq \mu_a^I$, we find (35) holds if

$$\frac{\hat{\lambda}_a^I}{\hat{\mu}_a^I} \triangleq k_b + \frac{\lambda_a^I}{\mu_a^I} + \frac{k_b\Theta_b}{\mu_a^I}.$$

By symmetry, we have a similar result for $y < 0$.

However, notice that because $\Theta_i > 0$, $\frac{\hat{\lambda}_i^I}{\hat{\mu}_i^I} > \frac{\hat{\hat{\lambda}}_i^I}{\hat{\hat{\mu}}_i^I}$ for $i = \{a, b\}$. Thus, rate of impatient units arriving in the hat-system is greater than that in the double-hat system, implying the former is more likely to be empty. Thus $0 < \hat{\hat{\mathbb{P}}}_0 < \hat{\mathbb{P}}_0$. By construction, $P_0 = \hat{P}_0$ and thus $\mathbb{P}_0 > 0$.

Thus we have shown that $-k_a - \frac{\lambda_b^I}{\mu_b^I} < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a} < \frac{\lambda_a^I}{\mu_a^I} + k_b$ implies that the density function takes the form of (7) and that $U_a > 0$, $U_b > 0$ and $0 < \mathbb{P}_0 < 1$. The other direction holds naturally since the parameters $\lambda_i, \mu_i, k_i, \lambda_i^I, \mu_i^I$, $i = \{a, b\}$ are all positive and we know that $\Theta_a > 0$ and $\Theta_b > 0$ if and only if $-k_a - \frac{\lambda_b^I}{\mu_b^I} < \frac{\lambda_b}{\mu_b} - \frac{\lambda_a}{\mu_a} < \frac{\lambda_a^I}{\mu_a^I} + k_b$. Q.E.D.

Proof of Proposition 2: Consider the system with no impatient customers, $\lambda_a^I = \lambda_b^I = \omega_a^I = \omega_b^I = 0$ (called the tilde system in the proof of Theorem 1). For $y > 0$, this case's solution satisfies (29) repeated here:

$$\tilde{r}(y)\tilde{f}(y) = \lambda_b \int_{-\infty}^{y} \exp\{\mu_b(z-y)\}\, \tilde{f}(z)dz + \tilde{\mathbb{P}}_0 \exp\{-\mu_b y\} - \lambda_a \int_{y}^{\infty} \exp\{-\mu_a(z-y)\}\, \tilde{f}(z)dz. \quad (36)$$

Now consider the original system. For $y > 0$, the level-crossing equation for the general model is

$$\begin{aligned} r(y)f(y) + \lambda_a^I \int_{y}^{\infty} \exp\left\{-\mu_a^I(z-y)\right\} f(z)dz \\ = \lambda_b \int_{-\infty}^{y} \exp\{\mu_b(z-y)\}\, f(z)dz + \mathbb{P}_0 \exp\{-\mu_b y\} - \lambda_a \int_{y}^{\infty} \exp\{-\mu_a(z-y)\}\, f(z)dz. \end{aligned} \quad (37)$$

By the assumption $-k_a - \lambda_b^I\omega_b^I < \lambda_b\omega_b - \lambda_a\omega_a < \lambda_a^I\omega_a^I + k_b$ and Theorem 1, $f(y)$ exists and is given by (7). Substituting $f(z) = U_b e^{-\Theta_b z}$ in the left hand side of (37) and integrating implies for $y > 0$,

$$\begin{aligned} k_b + \frac{\lambda_a^I}{\mu_a^I + \Theta_b} U_b e^{-\Theta_b y} \\ = \lambda_b \int_{-\infty}^{y} \exp\{\mu_b(z-y)\}\, f(z)dz + \mathbb{P}_0 \exp\{-\mu_b y\} - \lambda_a \int_{y}^{\infty} \exp\{-\mu_a(z-y)\}\, f(z)dz. \end{aligned} \quad (38)$$

Observe that because the right hand sides of (36) and (38) have the same form, letting $\tilde{k}_b = k_b + \frac{\lambda_a^I}{\mu_a^I + \Theta_b}$, then $\tilde{f}(y) = f(y)$ solves (36). By symmetry the same holds for the case of $y < 0$. Therefore $\tilde{\mathbb{P}}_0 = \mathbb{P}_0$. Observe that because $\Theta_i > 0$, $\bar{\bar{k}}_i = k_i + \lambda_j^I/\mu_j^I > \tilde{k}_i$ for $i \in \{a, b\}$. Thus the abandonment rate from the 'double-bar' system exceeds the abandonment rate from the 'tilde' system, and consequently $\bar{\bar{\mathbb{P}}}_0 > \mathbb{P}_0$. That $\hat{\hat{\mathbb{P}}} < \mathbb{P}_0$ was established in Theorem 1. □

Proof of Proposition 3: Let $\bar{M}_a(y) = \max\{n : \sum_{i=1}^{n} \bar{X}_{a,i} < y\}$. Observe $\bar{M}_a(y) + 1$ is a stopping time. Further, if $\bar{V}_{a,1}$, $\bar{V}_{a,2}$, ... are the inter-arrival times of the process $\{\bar{N}_a(t)\}$, then

$$\tau_b^c(y) = \sum_{i=1}^{\bar{M}_a(y)+1} \bar{V}_{a,i}.$$

Therefore by Wald's Equation,

$$\begin{aligned}\mathbb{E}[\tau_b^c(y)] &= \mathbb{E}\left[\sum_{i=1}^{\bar{M}_a(y)+1} \bar{V}_{a,i}\right] \\ &= \mathbb{E}\left[\bar{V}_{a,i}\right]\mathbb{E}[\bar{M}_a(y)+1] \\ &= E\left[\bar{V}_{a,i}\right](m(y)+1).\end{aligned}$$

where $m_a(y) = E[\bar{M}_a(y)]$. Observe $\mathbb{E}\left[\bar{V}_{a,i}\right] = 1/\bar{\lambda}_a$. From basic renewal theory (Karlin and Taylor 1975),

$$m(y) = G_a(y) + \int_0^y m(y-s)g_a(s)\,ds. \tag{39}$$

Substituting $g_a(s)$ and $G_a(y)$ into (39) and letting $z = y - s$,

$$m(y) = 1 - \frac{\lambda_a e^{-\mu_a y} + \lambda_a^I e^{-\mu_a^I y}}{\bar{\lambda}_a} + \int_0^y m(y)\left(\frac{\lambda_a}{\bar{\lambda}_a}\mu_a e^{-\mu_a(y-z)} + \frac{\lambda_a^I}{\bar{\lambda}_a}\mu_a^I e^{-\mu_a^I(y-z)}\right)dz.$$

The first and second derivatives are

$$m'(y) = \frac{\lambda_a\mu_a e^{-\mu_a y} + \lambda_a^I\mu_a^I e^{-\mu_a^I y}}{\bar{\lambda}_a} + \frac{\lambda_a\mu_a + \lambda_a^I\mu_a^I}{\bar{\lambda}_a}m(y) - \int_0^y m(z)\left(\frac{\lambda_a\mu_a^2 e^{-\mu_a(y-z)}}{\bar{\lambda}_a} + \frac{\lambda_a^I\mu_{\mathrm{aI}}^2 e^{-\mu_a^I(y-z)}}{\bar{\lambda}_a}\right)dz$$

$$\begin{aligned}m''(y) = &-\frac{\lambda_a\mu_a^2 e^{-\mu_a y} + \lambda_a^I\left(\mu_a^I\right)^2 e^{-\mu_a^I y}}{\bar{\lambda}_a} - \frac{\lambda_a\mu_a^2 + \lambda_a^I\left(\mu_a^I\right)^2}{\bar{\lambda}_a}m(y) + \frac{\lambda_a\mu_a + \lambda_a^I\mu_a^I}{\bar{\lambda}_a}m'(y) \\ &+ \int_0^y m(z)\left(\frac{\lambda_a\mu_a^3 e^{-\mu_a(y-z)}}{\bar{\lambda}_a} + \frac{\lambda_a^I\left(\mu_a^I\right)^3 e^{-\mu_a^I(y-z)}}{\bar{\lambda}_a}\right)dz\end{aligned}$$

Substituting the integrals into $m''(y)$ and simplifying yields

$$m''(y) = \mu_a\mu_a^I - \frac{\lambda_a^I\mu_a + \lambda_a\mu_a^I}{\bar{\lambda}_a}m'(y).$$

Solving this differential equation gives

$$m(y) = C_2 - C_1\frac{\bar{\lambda}_a}{\lambda_a^I\mu_a + \lambda_a\mu_a^I}e^{-\frac{y\left(\lambda_a^I\mu_a + \lambda_a\mu_a^I\right)}{\bar{\lambda}_a}} + \frac{\left(\bar{\lambda}_a\right)\mu_a\mu_a^I}{\lambda_a^I\mu_a + \lambda_a\mu_a^I}y,$$

where C_1 and C_2 are constants of integration solved for by using the boundary conditions: trivially $m(0) = 0$, and because the marginal number of arrivals required to empty the system as $y \to 0$ is inversely proportional to the arrival rate of units, $m'(0) = (\lambda_a\mu_a + \lambda_a^I\mu_a^I)/\bar{\lambda}_a$. Solving we find

$$C_1 = \frac{\lambda_a\lambda_a^I\left(\mu_a - \mu_a^I\right)^2}{\bar{\lambda}_a\left(\lambda_a^I\mu_a + \lambda_a\mu_a^I\right)},\ C_2 = \frac{\lambda_a\lambda_a^I\left(\mu_a - \mu_a^I\right)^2}{\left(\lambda_a^I\mu_a + \lambda_a\mu_a^I\right)^2}.$$

Consequently,

$$\begin{aligned}&\mathbb{E}[\tau_b^c(y)] = \frac{1}{\bar{\lambda}_a}(m(y)+1) \\ &= \frac{\left(\lambda_a^I\right)^2\mu_a^2 + \lambda_a^2\left(\mu_a^I\right)^2 + \lambda_a\lambda_a^I\left(\mu_a^2 + \left(\mu_a^I\right)^2\right)}{\bar{\lambda}_a\left(\lambda_a^I\mu_a + \lambda_a\mu_a^I\right)^2} + \frac{\mu_a\mu_a^I}{\lambda_a^I\mu_a + \lambda_a\mu_a^I}y - \frac{\lambda_a\lambda_a^I\left(\mu_a - \mu_a^I\right)^2}{\bar{\lambda}_a\left(\lambda_a^I\mu_a + \lambda_a\mu_a^I\right)^2}e^{-y\frac{\left(\lambda_a^I\mu_a + \lambda_a\mu_a^I\right)}{\bar{\lambda}_a}}.\end{aligned}$$

Substituting $\omega_a = 1/\mu_a$, $\omega_a^I = 1/\mu_a^I$ and simplifying gives the result. $\square$

Proof of Proposition 4: Letting $w = y - k_b t$, we can simplify (12) as

$$\begin{aligned}\mathbb{E}\left[\bar{T}_b^c(x,y)\right] =& \frac{1}{\lambda_a+\lambda_a^I} + H(x,y) \\ &+ \frac{1}{k_b}\int_0^y \int_0^w e^{-\left(\frac{\lambda_a+\lambda_a^I}{k_b}\right)(y-w)} \left(\lambda_a\mu_a e^{-\mu_a x} + \lambda_a^I\mu_a^I e^{-\mu_a^I x}\right) \mathbb{E}\left[\bar{T}_b^c(x,w-z)\right]\, dz\, dw.\end{aligned}$$

where

$$\begin{aligned}&H(x,y) \\ &= \frac{1}{(\lambda_a+\lambda_a^I)\left(\lambda_a^I\mu_a+\lambda_a\mu_a^I\right)^2} e^{-\frac{\lambda_a^2 y+\lambda_a^I\left(\lambda_a^I y+k_b\mu_a x\right)+\lambda_a\left(2\lambda_a^I y+k_b\mu_a^I x\right)}{k_b\left(\lambda_a+\lambda_a^I\right)}} \\ &\quad \Bigg(-\lambda_a\lambda_a^I\left(\mu_a-\mu_a^I\right)^2 \\ &\qquad + e^{\frac{\left(\lambda_a^I\mu_a+\lambda_a\mu_a^I\right)x}{\lambda_a+\lambda_a^I}} \left(\left(\lambda_a^I\right)^2\mu_a^2\mu_a^I x + \lambda_a^2\left(\mu_a^I\right)^2\mu_a x + \lambda_a\lambda_a^I\Big(\left(\mu_a^I\right)^2 + \mu_a\mu_a^I\left(-2+\mu_a^I x\right) + \mu_a^2\left(1+\mu_a^I x\right)\Big)\right)\Bigg) \\ &\quad + \frac{1}{\left(\lambda_a^I\mu_a+\lambda_a\mu_a^I\right)^2} \\ &\qquad \Bigg(\frac{\lambda_a\mu_a}{\lambda_a+\lambda_a^I-k_b\mu_a}\Bigg(\lambda_a(\mu_a^I)^2 x e^{-y\mu_a} - \lambda_a^I\mu_a^I e^{-y\mu_a} + \lambda_a^I\mu_a^I e^{-\frac{(y+x)\lambda_a^I\mu_a+\lambda_a\left(y\mu_a+x\mu_a^I\right)}{\lambda_a+\lambda_a^I}} \\ &\qquad\qquad + \lambda_a^I\mu_a e^{-y\mu_a} - \lambda_a^I\mu_a e^{-\frac{(y+x)\lambda_a^I\mu_a+\lambda_a\left(y\mu_a+x\mu_a^I\right)}{\lambda_a+\lambda_a^I}} + \lambda_a^I\mu_a\mu_a^I x e^{-y\mu_a}\Bigg) \\ &\qquad - \frac{\lambda_a\mu_a}{\lambda_a+\lambda_a^I-k_b\mu_a} e^{-\frac{y\left(\lambda_a+\lambda_a^I-k_b\mu_a\right)}{k_b}} \Bigg(\lambda_a\left(\mu_a^I\right)^2 x e^{-y\mu_a} - \lambda_a^I\mu_a^I e^{-y\mu_a} + \lambda_a^I\mu_a^I e^{-\frac{(y+x)\lambda_a^I\mu_a+\lambda_a\left(y\mu_a+x\mu_a^I\right)}{\lambda_a+\lambda_a^I}} \\ &\qquad\qquad + \lambda_a^I\mu_a e^{-y\mu_a} - \lambda_a^I\mu_a e^{-\frac{(y+x)\lambda_a^I\mu_a+\lambda_a\left(y\mu_a+x\mu_a^I\right)}{\lambda_a+\lambda_a^I}} + \lambda_a^I\mu_a\mu_a^I x e^{-y\mu_a}\Bigg) \\ &\qquad + \frac{\lambda_a^I\mu_a^I}{\lambda_a+\lambda_a^I-k_b\mu_a^I}\Bigg(\lambda_a^I\mu_a^2 x e^{-y\mu_a^I} + \lambda_a\mu_a^I e^{-y\mu_a^I} - \lambda_a\mu_a^I e^{-\frac{(y+x)\lambda_a\mu_a^I+\lambda_a^I\left(x\mu_a+y\mu_a^I\right)}{\lambda_a+\lambda_a^I}} \\ &\qquad\qquad - \lambda_a\mu_a e^{-y\mu_a^I} + \lambda_a\mu_a e^{-\frac{(y+x)\lambda_a\mu_a^I+\lambda_a^I\left(x\mu_a+y\mu_a^I\right)}{\lambda_a+\lambda_a^I}} + \lambda_a\mu_a e^{-y\mu_a^I} x\mu_a^I\Bigg) \\ &\qquad - \frac{\lambda_a^I\mu_a^I}{\lambda_a+\lambda_a^I-k_b\mu_a^I} e^{-\frac{y\left(\lambda_a+\lambda_a^I-k_b\mu_a^I\right)}{k_b}} \Bigg(\lambda_a^I\mu_a^2 x e^{-y\mu_a^I} + \lambda_a\mu_a^I e^{-y\mu_a^I} - \lambda_a\mu_a^I e^{-\frac{(y+x)\lambda_a\mu_a^I+\lambda_a^I\left(x\mu_a+y\mu_a^I\right)}{\lambda_a+\lambda_a^I}} \\ &\qquad\qquad - \lambda_a\mu_a e^{-y\mu_a^I} + \lambda_a\mu_a e^{-\frac{(y+x)\lambda_a\mu_a^I+\lambda_a^I\left(x\mu_a+y\mu_a^I\right)}{\lambda_a+\lambda_a^I}} + \lambda_a\mu_a\mu_a^I x e^{-y\mu_a^I}\Bigg)\Bigg)\end{aligned}$$

By taking a derivative with respect to y and equating the double integrals from the above equation and its derivative, we obtain an integro-differential equation of the form

$$\int_0^y \left(\lambda_a\mu_a e^{-\mu_a z} + \lambda_a^I\mu_a^I e^{-\mu_a^I z}\right) \mathbb{E}[\bar{T}_b^c(x,y-z)]\, dz$$

$$= \left(\lambda_a + \lambda_a^I\right) \mathbb{E}[\bar{T}_b^c(x,y)] + k_b \frac{\partial \mathbb{E}[\bar{T}_b^c(x,y)]}{\partial y} - \left(\lambda_a + \lambda_a^I\right) H(x,y) - k_b H_y(x,y)$$

where $H_y(x,y) = \partial H(x,y)/\partial y$. Letting $u = y - z$, we get

$$\int_0^y \left(\lambda_a \mu_a e^{-\mu_a(y-u)} + \lambda_a^I \mu_a^I e^{-\mu_a^I(y-u)}\right) \mathbb{E}[\bar{T}_b^c(x,u)]\, du$$
$$= \left(\lambda_a + \lambda_a^I\right) \mathbb{E}[\bar{T}_b^c(x,y)] + k_b \frac{\partial \mathbb{E}[\bar{T}_b^c(x,y)]}{\partial y} - \left(\lambda_a + \lambda_a^I\right) H(x,y) - k_b H_y(x,y)$$

We take the derivative with respect to y and equate the integral in the above equation and the same one found in the derivative. This removes one integral from the computation. Repeating this procedure, we obtain a homogeneous ordinary differential equation of the following form

$$\frac{\partial^3 \mathbb{E}[\bar{T}_b^c(x,y)]}{\partial y^3} + \frac{\lambda_a + \lambda_a^I + k_b\left(\mu_a + \mu_a^I\right)}{k_b} \frac{\partial^2 \mathbb{E}[\bar{T}_b^c(x,y)]}{\partial y^2}$$
$$+ \frac{\left(\lambda_a + k_b\mu_a\right)\mu_a^I + \lambda_a^I\left(\mu_a + 2\mu_a^I\right)}{k_b} \frac{\partial \mathbb{E}[\bar{T}_b^c(x,y)]}{\partial y} - \frac{\mu_a \mu_a^I}{k_b} = 0.$$

This third-order ODE can be solved yielding

$$\mathbb{E}[\bar{T}_b^c(x,y)] = C_0 + C_1 y - C_2 \frac{e^{-y\left(K_1 - \sqrt{K_2}\right)}}{K_1 - \sqrt{K_2}} - C_3 \frac{e^{-y\left(K_1 + \sqrt{K_2}\right)}}{K_1 + \sqrt{K_2}} + C_4(x)$$

where

$$K_1 = \left(\lambda_a + \lambda_a^I + k_b\mu_a + k_b\mu_a^I\right)/2k_b,$$
$$K_2 = \left(\lambda_a^2 + (\lambda_a^I)^2 + k_b^2\left(\mu_a - \mu_a^I\right)^2 - 2k_b\lambda_a^I\left(\mu_a + 3\mu_a^I\right) + 2\lambda_a\left(\lambda_a^I + k_b\left(\mu_a - \mu_a^I\right)\right)\right)/4k_b^2,$$

and $C_1 = \left(k_b + \frac{\lambda_a}{\mu_a} + \frac{\lambda_a^I}{\mu_a^I} + 2\frac{\lambda_a^I}{\mu_a}\right)^{-1}$. The other terms, C_0, C_2, C_3, and $C_4(x)$, are all independent of y and are obtained by solving three boundary conditions. Note that $K_1 \pm \sqrt{K_2}$ is positive for all non-negative parameter values.

The first boundary condition specifies that when a submitted order finds no units in the queue (i.e., $y = 0$) then the expected system time is equal to the clearing time:

$$\mathbb{E}[\bar{T}_b^c(x,0)] = \frac{\lambda_a\omega_a^2 + \lambda_a^I\left(\omega_a^I\right)^2}{\left(\lambda_a\omega_a + \lambda_a^I\omega_a^I\right)^2} + \frac{x}{\lambda_a\omega_a + \lambda_a^I\omega_a^I} - \frac{\lambda_a\lambda_a^I\left(\omega_a - \omega_a^I\right)^2}{\left(\lambda_a + \lambda_a^I\right)\left(\lambda_a\omega_a + \lambda_a^I\omega_a^I\right)^2} e^{-x\frac{\left(\lambda_a\omega_a + \lambda_a^I\omega_a^I\right)}{\left(\lambda_a + \lambda_a^I\right)\omega_a\omega_a^I}}.$$

The other conditions follow by observing that because the arrivals are compound Poisson, the marginal change in the system time is zero as the number of queued customers get small:

$$\lim_{y\to 0} \frac{\partial \mathbb{E}[\bar{T}_b^c(x,y)]}{\partial y} = 0 \quad \text{and} \quad \lim_{y\to 0} \frac{\partial^2 \mathbb{E}[\bar{T}_b^c(x,y)]}{\partial^2 y} = 0.$$

The result is that

$$
\begin{aligned}
C_0 &= \frac{\left(K_1^2 - K_2\right)\left(\frac{\lambda_a^I}{\left(\mu_a^I\right)^2} + \frac{\lambda_a}{\mu_a^2}\right)\left(k_b + \frac{\lambda_a}{\mu_a} + \frac{\lambda_a^I}{\mu_a^I} + 2\frac{\lambda_a^I}{\mu_a}\right) - 2K_1\left(\frac{\lambda_a^I}{\mu_a^I} + \frac{\lambda_a}{\mu_a}\right)^2}{\left(k_b + \frac{\lambda_a}{\mu_a} + \frac{\lambda_a^I}{\mu_a^I} + 2\frac{\lambda_a^I}{\mu_a}\right)\left(K_1^2 - K_2\right)\left(\frac{\lambda_a^I}{\mu_a^I} + \frac{\lambda_a}{\mu_a}\right)^2}, \\
C_2 &= -\frac{1}{2}C_1\frac{K_1 + \sqrt{K_2}}{\sqrt{K_2}}, \\
C_3 &= \frac{1}{2}C_1\frac{K_1 - \sqrt{K_2}}{\sqrt{K_2}}, \\
C_4(x) &= \frac{x}{\lambda_a\omega_a + \lambda_a^I\omega_a^I} - \frac{\lambda_a\lambda_a^I\left(\omega_a - \omega_a^I\right)^2}{\left(\lambda_a + \lambda_a^I\right)\left(\lambda_a\omega_a + \lambda_a^I\omega_a^I\right)^2}e^{-x\frac{\left(\lambda_a\omega_a + \lambda_a^I\omega_a^I\right)}{\left(\lambda_a + \lambda_a^I\right)\omega_a\omega_a^I}}.
\end{aligned}
$$

□

References

Karlin, S., Taylor, H. (1975). *A First Course in Stochastic Processes.* Academic Press.

Pachner, J. (1983). *Handbook of Numerical Analysis Applications with Programs for Engineers and Scientists.* McGraw-Hill, Inc.