mathematics

Article

Asymptotic Analysis for Systems with Deferred Abandonment

Katsunobu Sasanuma

College of Business, Stony Brook University, Stony Brook, NY 11794, USA; katsunobu.sasanuma@stonybrook.edu

Abstract: This short paper concerns the analysis of the M/M/k queueing system with customer abandonment. In this system, service managers provide a finite buffer space, which is a waiting area that prevents customers from abandoning the system. Abandonment of the system can occur from reneging (exiting from the queue while waiting), and/or balking (leaving the system without waiting). We derive an analytical expression to represent the impact of the buffer space capacity on the delay probability and the abandonment probability for a system with deferred abandonment. The result indicates the provision of the buffer space in a large system could only increase the delay probability while the abandonment probability remains unchanged. Despite the benevolent intentions of service managers, providing a buffer space may exacerbate the performance of larger systems.

Keywords: Markov chain; service systems; deferred abandonment; reneging; balking

check for updates

Citation: Sasanuma, K. Asymptotic Analysis for Systems with Deferred Abandonment. *Mathematics* **2021**, *9*, 2187. https://doi.org/10.3390/math9182187

Academic Editor: José Álvarez-García

Received: 10 July 2021
Accepted: 27 August 2021
Published: 7 September 2021

1. Introduction

Facility managers who operate large service systems such as call centers often face two conflicting goals when their systems are congested. The first goal is to reduce the probability that customers need to wait, which is represented by a delay probability P_Q. The second goal is to increase the number of customers they can serve (i.e., to increase the throughput of the system). Managers can increase their throughput by reducing the number of customers leaving the system due to reneging (exiting the queue while waiting) and/or balking (not entering the queue and walking away), or in other words, by reducing a customer abandonment probability P_{ab}. Facility managers typically prioritize one over the other, if both goals cannot be satisfied at the same time when a system is congested. In this paper, we provide a model that is helpful for managers to control the balance between these important performance indicators, P_Q and P_{ab}, when a simultaneous improvement is not possible.

Large service systems tend to exhibit high customer abandonment via reneging and/or balking. Such systems attract many prominent researchers and have been studied [1–4]. In recent years, many variations of systems with customer abandonment have also been studied: service slowdowns are incorporated to the system [5], availability of servers is time varying [6], and customers' patience depends on their individual service requirements [7]. Out of many models, the Erlang A model, an M/M/k+M queueing model with exponential reneging, is frequently used. The most important finding for the Erlang A model is the three asymptotic regimes describing the congestion properties of the system: Quality-and-Efficiency-Driven (QED), Quality-Driven (QD), and Efficiency-Driven (ED) regimes.

The square-root staffing rule in the QED regime plays an important role in the analysis of the Erlang A model [8,9]. The square-root staffing rule shows that by allocating a specific number of staff following the rule, facility managers can stably operate their systems. However, there are several elements commonly observed in reality, but not incorporated in the standard Erlang A model. First, the Erlang A model only considers reneging as a form of customer abandonment. In reality, customers not only renege, but may also balk when systems are heavily congested. If a customer knows there is going to be a significant amount of wait time, many customers would not want to enter the queue and will balk.

Second, facility managers often provide a buffer space between the queue and the service area to prevent abandonment. A buffer space is an area that customers can wait before proceeding to a service area. For example, in the lobby of a restaurant those at the front of the queue are in a buffer space, which does not incur reneging/balking; however, once the queue extends outside of the facility, abandonment is more likely. Another important example is the emergency department (ED), where many patients who need urgent medical attention randomly arrive. After being triaged and registered, patients often face a long wait in the buffer space. Patients who have been triaged and registered may not renege or balk, but those not yet triaged and waiting outside may renege at random times, or may balk and promptly leave for another hospital if the queue is long. In today's competitive market, the need for hospitals and urgent care to prevent balking or reneging is crucial from a fiscal perspective and from a medical perspective.

We propose a deferred abandonment model that represents a service facility with a buffer space, such as the emergency department and restaurants. Our deferred abandonment model is similar to the two-stage reneging queue discussed in [10], which represents a queue with two different reneging stages. Our model is different from the two-stage queue in three aspects: (1) Our model assumes a buffer space that does not allow customers/patients to abandon a queue, while the two-stage queue does not have such a buffer space, (2) our model allows either reneging or balking as an abandonment, while the two-stage queue only considers reneging, and (3) our model allows arrival and/or service rates to change when there's a queue. We study a system with deferred abandonment, derive the asymptotic formulae to represent its performance indicators, and analyze the impact of the buffer space on these measures. We show that despite the benevolent intentions of facility managers to improve the performance of their systems, providing a buffer space for customers could increase the waiting probability without improving the throughput when systems are large.

2. Deferred Abandonment Model

The deferred abandonment model represents a system where the first n ($\geq$0) customers in queue do not renege or do not incur state-dependent balking; abandonment is deferred by n as a result (see Figure 1). To analyze the properties of the deferred abandonment model, we split the system into sub-systems. We have three sub-chains: sub-chain 1 (an M/M/s/s queue: from states 0 to s), sub-chain 2 (a reneging/balking queue: from states t to infinity), and the buffer chain representing a buffer space, which is sub-chain 3 (an M/M/1/n queue: from states s to t), where states s and t are shared by neighboring sub-chains. We denote stationary probabilities of state k in the entire Markov chain and truncated sub-chain i as π_k and π_k^i, respectively.

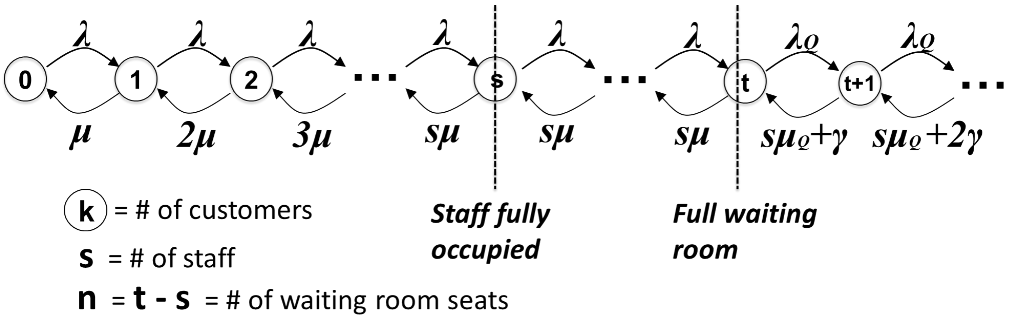

Figure 1. Deferred abandonment model (reneging system).

We consider two systems: reneging system and balking system. For the reneging system, we assume exponential reneging with rate $\gamma > 0$. For the balking system, we assume that the arrival rate drops by a linear balking rate $\delta > 0$ for each additional customer in queue. For either system, we allow changes in arrival and service rates when there is a queue that incurs abandonment, and denote these rates as $\lambda_Q = (1-\varepsilon)\lambda$ and $\mu_Q = (1+\tau)\mu$, respectively. (Thus, constant balking $\lambda - \lambda_Q = \varepsilon\lambda$ is incorporated in both (exponential) reneging and (linear) balking systems.) Let reneging/balking start at state $t(= s+n)$, where s is the number of staff. The birth and death coefficients of the Markov chain representing the deferred abandonment model are as follows:

1. For the reneging system, the total arrival rate and the total service rate at state k are
$$\lambda_k = \begin{cases} \lambda & 0 \le k < s \\ \lambda & s \le k < t \\ \lambda_Q & t \le k \end{cases} \text{ and } \mu_k = \begin{cases} k\mu & 1 \le k \le s \\ s\mu & s < k \le t \\ s\mu_Q + (k-t)\gamma & t < k \end{cases}.$$
2. For the balking system, the total arrival rate and the total service rate at state k are
$$\lambda_k = \begin{cases} \lambda & 0 \le k < s \\ \lambda & s \le k < t \\ (\lambda_Q - \delta\cdot(k-t))^+ & t \le k \end{cases} \text{ and } \mu_k = \begin{cases} k\mu & 1 \le k \le s \\ s\mu & s < k \le t \\ s\mu_Q & t < k \end{cases}.$$
Note: $(\cdot)^+$ denotes a positive part of a function.

Our deferred abandonment model can represent either the (exponential) reneging system ($\gamma > 0$) or the (linear) balking system ($\delta > 0$), both of which incorporate buffer spaces ($n \ge 0$), constant balking ($\varepsilon \ge 0$), and change of server speed (any τ). Our model is reduced to the original Erlang A/B/C models by choosing parameters appropriately. For example, if we set $\gamma > 0$ $(\delta = 0)$, $n = 0$ (thus $s = t$), and $\varepsilon = \tau = 0$, then our model becomes the standard Erlang A model (M/M/s+M queue). If $n = \infty$ when $s > \lambda/\mu$, our model approaches the Erlang C model (M/M/s queue). Finally, if $n = 0$ and $\varepsilon = 1$ (thus $\lambda_Q = 0$), our model becomes the Erlang B model (M/M/s/s queue).

Before concluding this section, we define several parameters to simplify the presentation of this paper. We define the resource requirement of sub-chain 1 as $R := \frac{\lambda}{\mu}$ and the resource requirement of sub-chain 2 as $R_Q := \frac{\lambda_Q}{\mu_Q}$. In this paper, we assume $R_Q \le R$ (because facility managers always try to reduce the level of congestion when the system is congested). We define linear and square-root staffing coefficients as $a := \frac{s-R}{R}$ (≥ -1) and $c := \frac{s-R}{\sqrt{R}}$, respectively. Server utilization is defined as $\rho := \frac{\lambda}{s\mu}$. We also define $a_Q := \frac{s-R_Q}{R} = a + \frac{R-R_Q}{R} = a + \frac{\varepsilon+\tau}{1+\tau}$. Other symbols are defined as needed throughout the paper.

3. Analysis of the Deferred Abandonment Model

We define the following performance indicators: P_Q is the probability that a customer enters a queue with abandonment (reneging or balking) (i.e., $P_Q := \sum_{k=t}^{\infty} \pi_k$); P_{ab} is the probability that customers abandon a queue via constant balking or exponential reneging (for the reneging system), or abandon a queue via constant balking or linear balking (for the balking system); P_W is the probability that an arriving customer sits in one of the n seats in the buffer space (i.e., $P_W := \sum_{k=s}^{t-1} \pi_k$). (Note: $P_W = 0$ when $n = 0$.) We define the delay probability for this system as $P_{Q+} := \sum_{k=s}^{\infty} \pi_k = P_W + P_Q$. We represent these performance indicators by the blocking probabilities of three truncated sub-chains: $\pi_s^1, \pi_t^2, \pi_s^3$, and π_t^3. For this purpose, we utilize Kelly's property (Corollary 1.10. in [11]) that holds for a reversible Markov chain. (Note: The extension of the property to more general Markov chains is discussed in [12].) Since our deferred abandonment model is a reversible Markov chain, the entire Markov chain and its truncated sub-chains satisfy Kelly's property:

Lemma 1 (Kelly's property)**.** *Suppose that a Markov chain is reversible. Let P_i be the probability that we observe a state in sub-chain i. $P_i = \pi_k/\pi_k^i$ holds for any sub-chain i and any state k in sub-chain i.*

The proof of Lemma 1 is omitted (see [11] or [12] if interested.) Lemma 1 essentially states that a truncated sub-chain has the same stationary distribution as the distribution of the entire Markov chain given in the sub-chain. Lemma 1 allows us to work on the individual truncated sub-chain rather than on the entire Markov chain, simplifying the analysis of Markov chain models substantially. Using Lemma 1 repeatedly, we can derive the exact relationships among performance indicators and blocking probabilities, which are summarized in Lemma 2.

Lemma 2. *The following structural representation holds for the deferred abandonment model:*

$$\frac{1}{\pi_s} = \frac{1}{\pi_s^1} + \frac{1}{\pi_s^3/\pi_t^3} \cdot \left(\frac{1}{\pi_t^2} - 1\right) + \left(\frac{1}{\pi_s^3} - 1\right),$$

$$\frac{P_W}{\pi_s} = \frac{1}{\pi_s^3/\pi_t^3} \cdot \left(\frac{1}{\pi_t^3} - 1\right),$$

$$\frac{P_Q}{\pi_s} = \frac{1}{\pi_s^3/\pi_t^3} \cdot \frac{1}{\pi_t^2},$$

$$\frac{P_{ab}}{\pi_s} = \frac{1}{\pi_s^3/\pi_t^3} \cdot \frac{P_{ab}^2}{\pi_t^2},$$

where

$$\frac{P_{ab}^2}{\pi_t^2} = 1 + p \cdot \left(\frac{1}{\pi_t^2} - 1\right)$$

and

$$p = 1 - \frac{s\mu_Q}{\lambda} = 1 - (1+\tau)(a+1).$$

Proof of Lemma 2. We denote the sub-chain that combines sub-chain 3 and sub-chain 2 as sub-chain 2+. By viewing chain 2+ as the entire chain and chain 3 as a sub-chain of chain 2+, we can apply Lemma 1 to states s and t that belong to sub-chain 2+, and obtain the relationship $\pi_s^{2+}/\pi_s^3 = \pi_t^{2+}/\pi_t^3$, from which we obtain $\pi_s^{2+} = (\pi_s^3/\pi_t^3)\pi_t^{2+}$. Additionally, using Lemma 1, we can show

$$1 = P_1 + P_{2+} - \pi_s = \frac{\pi_s}{\pi_s^1} + \frac{\pi_s}{\pi_s^{2+}} - \pi_s,$$

from which we obtain

$$\frac{1}{\pi_s} = \frac{1}{\pi_s^1} + \frac{1}{\pi_s^{2+}} - 1,$$

and likewise,

$$\frac{1}{\pi_t^{2+}} = \frac{1}{\pi_t^2} + \frac{1}{\pi_t^3} - 1.$$

Combining these results, we can derive $1/\pi_s$:

$$\frac{1}{\pi_s} = \frac{1}{\pi_s^1} + \frac{1}{\pi_s^{2+}} - 1 = \frac{1}{\pi_s^1} + \frac{1}{\pi_s^3/\pi_t^3} \cdot \frac{1}{\pi_t^{2+}} - 1 = \frac{1}{\pi_s^1} + \frac{1}{\pi_s^3/\pi_t^3} \cdot \left(\frac{1}{\pi_t^2} + \frac{1}{\pi_t^3} - 1\right) - 1$$
$$= \frac{1}{\pi_s^1} + \frac{1}{\pi_s^3/\pi_t^3} \cdot \left(\frac{1}{\pi_t^2} - 1\right) + \left(\frac{1}{\pi_s^3} - 1\right).$$

We can also derive P_W and P_Q using Lemma 1:

$$P_W = \sum_{k=s}^{t-1} \pi_k = \sum_{k=s}^{t} \pi_k - \pi_t = P_3 - \pi_t = \frac{\pi_t}{\pi_t^3} - \pi_t$$
$$= \pi_t \cdot \left(\frac{1}{\pi_t^3} - 1\right) = \pi_s \cdot \frac{1}{\pi_s^3/\pi_t^3} \cdot \left(\frac{1}{\pi_t^3} - 1\right)$$

and

$$P_Q = \sum_{k=t}^{\infty} \pi_k = P_2 = \frac{\pi_t}{\pi_t^2} = \pi_s \cdot \frac{1}{\pi_s^3/\pi_t^3} \cdot \frac{1}{\pi_t^2}.$$

Finally, notice that the abandonment occurs only at the reneging/balking sub-chain (i.e., sub-chain 2) and the probability of abandonment given sub-chain 2 is $p = \frac{\lambda - s\mu_Q}{\lambda} = 1 - \frac{s\mu_Q}{\lambda}$. Thus, using Lemma 1 again, we obtain

$$P_{ab} = P_{ab}^2 \cdot P_2 = P_{ab}^2 \cdot \frac{\pi_t}{\pi_t^2} = \pi_s \cdot \frac{1}{\pi_s^3/\pi_t^3} \cdot \frac{P_{ab}^2}{\pi_t^2},$$

where

$$\frac{P_{ab}^2}{\pi_t^2} = \frac{\pi_t^2 + p \cdot (1 - \pi_t^2)}{\pi_t^2} = 1 + p \cdot \left(\frac{1}{\pi_t^2} - 1 \right).$$

□

Remark 1. *By plugging the exact, approximate, or asymptotic limit of blocking probabilities $\pi_s^1, \pi_t^2, \pi_s^3$, and π_t^3 into Lemma 2, we can derive the exact, approximate, or asymptotic limit of performance indicators for the deferred abandonment model, respectively.*

For the rest of this section, we show that the two important indicators, the delay probability P_{Q+} and the abandonment probability P_{ab}, exhibit a trade-off relationship when the number of buffer spaces n changes. For this purpose, we denote performance indicators of the deferred abandonment model as an explicit function of n: $\pi_s(n)$, $P_W(n)$, $P_Q(n)$, $P_{Q+}(n)$, and $P_{ab}(n)$. When $n = 0$, sub-chain 3 is reduced to a single state s; thus, $s = t$, $\pi_s^3 = \pi_t^3 = 1$, $P_W(0) = 0$, and $P_{Q+}(0) = P_Q(0)$ hold. To simplify the representation, we introduce functions $\widetilde{P}_Q$ and $\widetilde{P}_{ab}$ that represent delay and abandonment probabilities for the Markov chain model which comprises two of the three sub-chains in the deferred abandonment model: sub-chains 1 and 2, sub-chains 1 and 3, or sub-chains 3 and 2. (Note: Abandonment probability is defined properly only when the right sub-chain is sub-chain 2.) When $n = 0$, the model is composed of sub-chains 1 and 2, and thus,

$$P_{Q+}(0) = \frac{\frac{1}{\pi_s^2}}{\frac{1}{\pi_s^1} + \frac{1}{\pi_s^2} - 1} =: \widetilde{P}_Q\left(\frac{1}{\pi_s^1}, \frac{1}{\pi_s^2}\right) \text{ and } P_{ab}(0) = \frac{1 + p \cdot (\frac{1}{\pi_s^2} - 1)}{\frac{1}{\pi_s^1} + \frac{1}{\pi_s^2} - 1} =: \widetilde{P}_{ab}\left(\frac{1}{\pi_s^1}, \frac{1}{\pi_s^2}\right).$$

To prove the trade-off relationship between $P_{Q+}(n)$ and $P_{ab}(n)$, the following lemma is necessary.

Lemma 3. *For any $a \in (-1, \infty)$ (i.e., $\forall s > 0$) for the deferred abandonment model, the following inequalities hold:*

$$\frac{a}{\pi_s^1} + 1 > 0 \text{ and } \frac{a}{\pi_t^2} - a - 1 < 0.$$

Proof of Lemma 3. For the first relationship, consider the $-1 < a < 0$ case only, since the relationship is trivial when $a \geq 0$. Assuming $a < 0$, the first relationship is equivalent to $1/\pi_s^1 < -1/a$. Note that from the definitions of arrival/service rates for the left sub-chain, $\forall i \in \{0, 1, 2, \cdots, s-1\}$, $\mu_{s-i} \leq s\mu$, $\lambda_{s-i-1} = \lambda$, and hence, $\mu_{s-i}/\lambda_{s-i-1} \leq s\mu/\lambda = \rho^{-1} = a + 1$, where the equality holds only at $i = 0$. Notice also that $0 < \rho^{-1} < 1$ when $-1 < a < 0$. Using these properties and the exact result, we can derive the first relationship:

$$\frac{1}{\pi_s^1} = 1 + \sum_{k=1}^{s} \prod_{i=0}^{k-1} \frac{\mu_{s-i}}{\lambda_{s-i-1}} \leq 1 + \sum_{k=1}^{s} \rho^{-k} < 1 + \sum_{k=1}^{\infty} \rho^{-k} = 1 + \frac{\rho^{-1}}{1 - \rho^{-1}} = \frac{1}{1 - \rho^{-1}} = -\frac{1}{a}.$$

Next, for the second relationship, consider the case $a > 0$ only, since the relationship is trivial when $a \leq 0$. Assuming $a > 0$, the second relationship is equivalent to $1/\pi_s^2 < 1 + (1/a)$. Note that from the definitions of arrival/service rates for the right (reneging/balking) part of the Markov chain, $\forall i \in \{0, 1, \cdots\}$, $\lambda_{t+i} \leq \lambda_Q$, $\mu_{t+i+1} \geq s\mu_Q$, and hence, $\lambda_{t+i}/\mu_{t+i+1} \leq \lambda_Q/(s\mu_Q) \leq \lambda/(s\mu) = \rho = 1/(a+1)$, where the equality would hold when both $\gamma = \delta = 0$ and $\varepsilon + \tau = 0$ hold. (Note that the equality does not hold since we assume that reneging/balking must exist.) Notice also that $0 < \rho < 1$ when $a > 0$. Using these properties and the exact result, we can derive the second relationship:

$$\frac{1}{\pi_t^2} = 1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_{t+i}}{\mu_{t+i+1}} < 1 + \sum_{k=1}^{\infty} \rho^k = 1 + \frac{\rho}{1-\rho} = \frac{1}{1-\rho} = 1 + \frac{1}{a}.$$

□

Using Lemma 3 and performance indicators $\widetilde{P}_Q$ and $\widetilde{P}_{ab}$ for the two sub-chain model, we can now prove the trade-off relationship between $P_{Q+}(n)$ and $P_{ab}(n)$.

Proposition 1. *For the deferred abandonment system, $P_{Q+}(n)$ and $P_{ab}(n)$ show the trade-off relationship as n changes.*

1. *$P_{Q+}(n)$ monotonically increases as n increases:*
 The lower bound is $\widetilde{P}_Q(1/\pi_s^1, 1/\pi_s^2)$ at $n = 0$ and the upper bound is either $\widetilde{P}_Q(1/\pi_s^1, 1 + 1/a)$ (when $a > 0$) or 1 (when $a \leq 0$) at $n \to \infty$.
2. *$P_{ab}(n)$ monotonically decreases as n increases:*
 The upper bound is $\widetilde{P}_{ab}(1/\pi_s^1, 1/\pi_s^2)$ at $n = 0$ and the lower bound is either 0 (when $a \geq 0$) or $\widetilde{P}_{ab}(-1/a, 1/\pi_t^2)$ (when $a < 0$) at $n \to \infty$.

Remark 2. *In Proposition 1, performance indicators of the two sub-chain model (i.e., $\widetilde{P}_Q$ and $\widetilde{P}_{ab}$) show up; this is explained intuitively as follows: When sub-chain 3 (middle sub-chain) does not exist ($n = 0$), the deferred abandonment model with three sub-chains (left, middle, and right) becomes the model with only sub-chains 1 and 2 (left and right). Likewise, when sub-chain 3 is infinitely large ($n \to \infty$), the deferred abandonment model becomes equivalent to the model with either sub-chains 1 and 3 (left and middle) if $\rho < 1$, or sub-chains 3 and 2 (middle and right) if $\rho > 1$. Since the upper/lower bounds of $P_{Q+}(n)$ and $P_{ab}(n)$ are obtained at either $n = 0$ or $n \to \infty$, the properties of the deferred abandonment model can be described by the properties of the two sub-chain model. Note that the following properties hold for sub-chain 3: $1/\pi_s^3 = 1/(1-\rho) = 1 + 1/a$ at $n \to \infty$ when $a > 0$ (thus $s > R$ and $\rho < 1$) and $1/\pi_t^3 = 1/(1-\rho^{-1}) = -1/a$ at $n \to \infty$ when $a < 0$ (thus $s < R$ and $\rho > 1$).*

Proposition 1 indicates that the trade-off relationship exists between $P_{Q+}(n)$ and $P_{ab}(n)$. If we provide more seats for customers in a buffer space in an abandonment system, we are able to reduce the number of customers abandoning the system (i.e., reduce $P_{ab}(n)$) at the cost of higher delay probability for arriving customers (i.e., increase $P_{Q+}(n)$). For the remainder of this section, we show the proof of Proposition 1.

Proof of Proposition 1. For the deferred abandonment model, we assume that sub-chain 3 is an M/M/1/n queue with $\rho = \lambda/(s\mu)$. We consider two possible cases: (1) $\rho = 1$ and (2) $\rho \neq 1$. For the second case, we utilize Lemma 3.
(1) $\rho = 1$: In this case, sub-chain 3 (an M/M/1/n queue with $\rho = 1$) satisfies

$$\frac{1}{\pi_s^3/\pi_t^3} = 1 \text{ and } \frac{1}{\pi_s^3} = n + 1.$$

Combined with Lemma 2, we obtain

$$\frac{1}{\pi_s(n)} = \frac{1}{\pi_s^1} - 1 + \frac{1}{\pi_t^2} + n, \frac{P_W(n)}{\pi_s(n)} = n, \frac{P_Q(n)}{\pi_s(n)} = \frac{1}{\pi_t^2}, \frac{P_{ab}(n)}{\pi_s(n)} = 1 + p \cdot \left(\frac{1}{\pi_t^2} - 1\right),$$

from which we obtain

$$P_{Q+}(n) = P_W(n) + P_Q(n) = \frac{\frac{1}{\pi_t^2} + n}{\frac{1}{\pi_s^1} - 1 + \frac{1}{\pi_t^2} + n} \text{ and } P_{ab} = \frac{1 + p \cdot \left(\frac{1}{\pi_t^2} - 1\right)}{\frac{1}{\pi_s^1} - 1 + \frac{1}{\pi_t^2} + n}.$$

This result obviously satisfies the properties shown in Proposition 1. $P_{Q+}(n)$ monotonically increases as n increases; the lower bound is $P_Q(1/\pi_s^1, 1/\pi_s^2)$ at $n = 0$; the upper bound is 1 at $n \to \infty$. $P_{ab}(n)$ monotonically decreases as n increases; the upper bound is $P_{ab}(1/\pi_s^1, 1/\pi_s^2)$ at $n = 0$; the lower bound is 0 at $n \to \infty$.

(2) $\rho \neq 1$: In this case, sub-chain 3 (an M/M/1/n queue with $\rho \neq 1$) satisfies

$$\frac{1}{\pi_s^3/\pi_t^3} = \rho^n \text{ and } \frac{1}{\pi_s^3} = \frac{1 - \rho^{n+1}}{1 - \rho}.$$

To obtain a representation in terms of a linear coefficient a, we further rewrite these equations using the relationships, $\rho = \lambda/(s\mu) = 1/(a+1)$, $\rho/(1-\rho) = 1/a$, and $1/(1-\rho) = 1 + (1/a)$, and we obtain

$$\frac{1}{\pi_s^3} - 1 = (1 - \rho^n)\frac{1}{a} \text{ and } \frac{1}{\pi_t^3} - 1 = \frac{1 - \rho^n}{\rho^n}\left(1 + \frac{1}{a}\right).$$

Combined with Lemma 2, we obtain

$$\frac{1}{\pi_s(n)} = \frac{1}{\pi_s^1} + \rho^n \cdot \left(\frac{1}{\pi_t^2} - 1\right) + (1 - \rho^n)\frac{1}{a},$$

$$\frac{P_W(n)}{\pi_s(n)} = (1 - \rho^n)\left(1 + \frac{1}{a}\right), \frac{P_Q(n)}{\pi_s(n)} = \frac{\rho^n}{\pi_t^2}, \frac{P_{ab}(n)}{\pi_s(n)} = \rho^n \cdot \left[1 + p \cdot \left(\frac{1}{\pi_t^2} - 1\right)\right],$$

from which we obtain

$$P_{Q+}(n) = \frac{\xi(n)}{\frac{1}{\pi_s^1} + \xi(n) - 1} = \widetilde{P}_Q\left(\frac{1}{\pi_s^1}, \xi(n)\right), \text{ where } \xi(n) := 1 + \frac{1}{a} + \rho^n \cdot \left(\frac{a}{\pi_t^2} - a - 1\right)\frac{1}{a},$$

and

$$P_{ab}(n) = \frac{1 + p \cdot \left(\frac{1}{\pi_t^2} - 1\right)}{\eta(n) + \frac{1}{\pi_t^2} - 1} = \widetilde{P}_{ab}\left(\eta(n), \frac{1}{\pi_t^2}\right), \text{ where } \eta(n) := -\frac{1}{a} + \rho^{-n} \cdot \left(\frac{a}{\pi_s^1} + 1\right)\frac{1}{a}.$$

As we expect, when $n = 0$, we retrieve $\widetilde{P}_Q(1/\pi_s^1, 1/\pi_s^2)$ and $\widetilde{P}_{ab}(1/\pi_s^1, 1/\pi_s^2)$.

To prove that $P_{Q+}(n)$ increases monotonically as n increases, notice that $\xi(n)$ is a positive increasing function of n. Regardless of the value of a (note: $\rho = 1/(a+1) < 1$ holds for $a > 0$, and $\rho = 1/(a+1) > 1$ holds for $-1 < a < 0$), using Lemma 3, $\xi(n)$ is a monotonically increasing function of n which satisfies

$$\xi(n) = 1 + \frac{1}{a} + \rho^n \cdot \left(\frac{a}{\pi_t^2} - a - 1\right)\frac{1}{a} \geq \frac{1}{\pi_t^2}(> 0),$$

where the equality holds at $n = 0$. We can conclude that $P_{Q+}(n)$ is an increasing function of n where the lower bound is $\widetilde{P}_Q(1/\pi_s^1, 1/\pi_s^2)$ obtained at $n = 0$, and the upper bound is either $\widetilde{P}_Q(1/\pi_s^1, 1 + 1/a)$ (when $a > 0$) or 1 (when $a < 0$) at $n \to \infty$.

To prove that $P_{ab}(n)$ decreases monotonically as n increases, notice that $\eta(n)$ is a positive increasing function of n. Regardless of the value of a, using Lemma 3, $\eta(n)$ is a monotonically increasing function of n which satisfies

$$\eta(n) = -\frac{1}{a} + \rho^{-n} \cdot \left(\frac{a}{\pi_s^1} + 1\right)\frac{1}{a} \geq \frac{1}{\pi_s^1}(> 0),$$

where the equality holds at $n = 0$. We can conclude that $P_{ab}(n)$ is a decreasing function of n where the the upper bound is $\widetilde{P}_{ab}(1/\pi_s^1, 1/\pi_s^2)$ obtained at $n = 0$, and the lower bound is either 0 (when $a > 0$) or $\widetilde{P}_{ab}(-1/a, 1/\pi_t^2)$ (when $a < 0$) at $n \to \infty$. □

4. Asymptotic Representation of Systems with Deferred Abandonment

In Proposition 1, we observe that there exists a trade-off relationship between the delay probability and the abandonment probability when the size of the buffer space changes. This is true for any systems with smaller (finite) resource requirement R. However, what if when the system grows large? In fact, many systems exhibit larger R compared to the number of buffer spaces n. In this section, we analyze the asymptotic limit of larger systems, obtain useful linear/square-root staffing rules, and discuss the trade-off relationship for larger systems. To find the asymptotic limit of performance indicators, all we need to know is the asymptotic limit of blocking probabilities for sub-chains 1 and 2 since sub-chain 3 is only affected by n and not by R.

To represent asymptotic results, we first define the necessary parameters in Table 1. For simplicity, we use square-root coefficients $c = \frac{s-R}{\sqrt{R}}$, $c' = \frac{s'-R'}{\sqrt{R'}}$, $c'' = \frac{s''-R''}{\sqrt{R''}}$, and linear coefficients $a = \frac{s-R}{R}$, $a' = \frac{s'-R'}{R'}$, $a'' = \frac{s''-R''}{R''}$, which correspond to sub-chain 1 (M/M/s/s), sub-chain 2 (reneging), and sub-chain 2 (balking), respectively. Following the normal approximation described in [10], we obtain Lemma 4: Blocking probabilities of sub-chains are approximated by the hazard function for the standard normal distribution $h(\cdot)$ and the continuity correction terms $\Delta = \frac{0.5}{\sqrt{R}}$, $\Delta' = \frac{0.5}{\sqrt{R'}}$, and $\Delta'' = \frac{0.5}{\sqrt{R''}}$.

Table 1. Definition of parameters.

Sub-Chain	Square-Root Coef	Linear Coef	Resource Requirement	Staffing Level
M/M/s/s	$c\ = a\sqrt{R}$	$a\ = \dfrac{c}{\sqrt{R}}$	$R = \dfrac{\lambda}{\mu}\ = \left(\dfrac{c}{a}\right)^2$	s
Reneging	$c' = \dfrac{a_Q}{a}\sqrt{\dfrac{1+\tau}{1-\varepsilon}}\sqrt{\dfrac{\mu_Q}{\gamma}} \cdot c$	$a' = \dfrac{1+\tau}{1-\varepsilon} \cdot a_Q$	$R' = \dfrac{\lambda_Q}{\gamma} = \dfrac{(1-\varepsilon)\mu}{\gamma} \cdot R$	$s' = \dfrac{s\mu_Q}{\gamma} = \dfrac{\mu_Q}{\gamma} \cdot s$
Balking	$c'' = -\dfrac{a_Q}{a}\sqrt{\dfrac{\mu_Q}{(a+1)\delta}} \cdot c$	$a'' = -\dfrac{a_Q}{a+1}$	$R'' = \dfrac{s\mu_Q}{\delta} = \dfrac{(a+1)\mu_Q}{\delta} \cdot R$	$s'' = \dfrac{\lambda_Q}{\delta} = \dfrac{1-\varepsilon}{1+\tau} \cdot \dfrac{\mu_Q}{(a+1)\delta} \cdot s$

Lemma 4. *Blocking probabilities of sub-chain 1 (M/M/s/s sub-chain) and sub-chain 2 (reneging or balking sub-chain) are approximately represented using the standard normal hazard function $h(\cdot)$:*

$$\frac{1}{\pi_s^1} \approx \frac{c}{a \cdot h(-c-\Delta)}, \frac{1}{\pi_t^2} \approx 1 + \frac{c'}{a' \cdot h(c'+\Delta')} \textit{ for reneging sub-chain,}$$

$$\textit{and } \frac{1}{\pi_t^2} \approx \frac{c''}{a'' \cdot h(-c''-\Delta'')} \textit{ for balking sub-chain.}$$

We omit the proof of Lemma 4, as it is almost identical to that given in [10]. The key idea of this approximation is to represent blocking probabilities of sub-chains by the Poisson representation, and approximate them by the standard normal representation. The averages of the three Poisson distributions for sub-chains 1, 2 (reneging), and 2 (balking) when blocking probabilities are represented by the Poisson random variables are R, R' and R'', and the continuity correction terms when converting the discrete Poisson distribution to the continuous standard normal distribution are Δ, Δ', and Δ'', respectively. The Poisson-to-normal approximation is elementary, but highly accurate when the average of the Poisson distribution is around 10 or more, and the approximation becomes exact when the average goes to infinity. Thus, Lemma 4 accurately represents the blocking probabilities of all sub-chains as $R \to \infty$ (which leads to $R', R'' \to \infty$ as well). We now obtain two asymptotic limits of blocking probabilities: (i) linear staffing rule (when $R_Q < R$); and (ii) square-root staffing rule (when $R_Q = R$).

Lemma 5 (Linear staffing asymptotic regime)**.** *Let $s = R + aR$ (or $a = \frac{s-R}{R}$) and take the limit of large R with fixed a. Then*

1. *For sub-chain 1 (M/M/s/s sub-chain)*

$$\frac{1}{\pi_s^1} \to \begin{cases} -\frac{1}{a} & s = R + aR < R\,(\textit{when } a < 0), \\ \infty & s = R + aR > R\,(\textit{when } a > 0). \end{cases}$$

2. *For sub-chain 2 (either reneging or balking sub-chain)*

$$\frac{1}{\pi_t^2} \to \begin{cases} \infty & s = R + aR < R_Q\,(\textit{when } a_Q < 0), \\ \frac{a+1}{a_Q} & s = R + aR > R_Q\,(\textit{when } a_Q > 0). \end{cases}$$

Proof of Lemma 5. We use the properties of the standard normal hazard function in this proof: $x/h(x) \to 1$ as $x \to \infty$ and $x/h(x) \to -\infty$ as $x \to -\infty$. Now, consider taking the limit of large R while fixing $a = \frac{s-R}{R}$. For sub-chain 1, if $a < 0$ (i.e., $s < R$), then $c = a\sqrt{R} \to -\infty$, and thus, $\frac{1}{\pi_s^1} \to -\frac{1}{a}$ as $R \to \infty$; and if $a > 0$ (i.e., $s > R$), then $c = a\sqrt{R} \to \infty$, and thus, $\frac{1}{\pi_s^1} \to \infty$ as $R \to \infty$.

For sub-chain 2, if $a_Q < 0$ (i.e., $s < R_Q$), then $a' < 0$ and $a'' > 0$, both of which are fixed. Thus, $c' = a'\sqrt{R'} \to -\infty$ and $c'' = a''\sqrt{R''} \to +\infty$, leading to $\frac{1}{\pi_t^2} \to \infty$ as $R \to \infty$. (Note that $R', R'' \to \infty$ as $R \to \infty$; see Table 1) Likewise, if $a_Q > 0$ (i.e., $s > R_Q$), then $a' > 0$ and $a'' < 0$. Thus, $c' = a'\sqrt{R'} \to +\infty$ and $c'' = a''\sqrt{R''} \to -\infty$, leading to $\frac{1}{\pi_t^2} \to 1 + \frac{1}{a'}$ (or $-\frac{1}{a''}$) $= \frac{a+1}{a_Q}$ as $R \to \infty$. □

Lemma 6 (Square-root staffing asymptotic regime)**.** *Assume $R_Q = R$ (thus, $\epsilon + \tau = 0$). Let $s = R + c\sqrt{R}$ (or $c = \frac{s-R}{\sqrt{R}}$) and $a = \frac{c}{\sqrt{R}}$. By taking the limit of large R with fixed c, we obtain:*

1. *For sub-chain 1 (M/M/s/s sub-chain)*

$$\frac{1}{\pi_s^1} \to +\infty \quad \textit{and} \quad \frac{a}{\pi_s^1} \to \frac{c}{h(-c)}.$$

2. *For sub-chain 2 (either reneging or balking sub-chain)*

$$\frac{1}{\pi_t^2} \to +\infty \quad \textit{and} \quad \frac{a}{\pi_t^2} \to \frac{\sqrt{\frac{\mu_Q}{\theta}} \cdot c}{h\left(\frac{\mu_Q}{\theta} \cdot c\right)}.$$

where $\theta = \gamma$ (or δ) for reneging (or balking) sub-chain.

Proof of Lemma 6. We take the limit of large R while fixing $c = \frac{s-R}{\sqrt{R}}$. Thus, $a = \frac{c}{\sqrt{R}} \to 0$. Additionally, using Table 1 and the assumption $R_Q = R$ (thus $\epsilon + \tau = 0$), we obtain $c' = \sqrt{\frac{\mu_Q}{\gamma}} \cdot c$, $a' = a_Q = a$, $c'' = -\sqrt{\frac{\mu_Q}{(a+1)\gamma}} \cdot c \to -\sqrt{\frac{\mu_Q}{\gamma}} \cdot c$, and $a'' = -\frac{a}{a+1}$. Finally, all continuity correction terms become negligible in the limit of large R: $\Delta, \Delta', \Delta'' \to 0$. Combining these results with Lemma 4, we obtain the result of Lemma 6. □

Combining Lemmas 2, 5 and 6 with the assumption $R_Q \leq R$, we obtain Proposition 2. Proposition 2 describes the asymptotic representation of performance indicators given n (or more specifically, we take the limit of large R while fixing n). Notice that the $n = \epsilon = \tau = 0$ case (thus, $R_Q = R$ holds) corresponds to the asymptotic formulae for the Erlang A model. We define a function needed for the square-root staffing rule: $\phi(c) := \frac{\sqrt{\mu_Q/\theta}}{h(\sqrt{\mu_Q/\theta}\cdot c)}\left(\frac{1}{h(-c)} + \frac{\sqrt{\mu_Q/\theta}}{h(\sqrt{\mu_Q/\theta}\cdot c)}\right)^{-1}$, where $\theta = \gamma$ (or δ) for a reneging (or balking) system.

Proposition 2. *We consider three asymptotic regimes for the deferred abandonment model:*

1. *ED asymptotic regime: We take the limit of large R while fixing the linear coefficient a that satisfies $s = R + aR < R_Q$ and obtain*

$$P_{Q+}(n) \to 1 \ \textit{and} \ P_{ab}(n) \to p.$$

2. *QD asymptotic regime: We take the limit of large R while fixing the linear coefficient a that satisfies $s = R + aR > R$ and obtain*

$$P_{Q+}(n) \to 0 \ \textit{and} \ P_{ab}(n) \to 0.$$

3. *QED asymptotic regime: There are two QED asymptotic regimes.*
   (a) *(Linear staffing rule) When $R_Q < R$, we take the limit of large R while fixing the linear coefficient a that satisfies $R_Q < s = R + aR < R$ and obtain*

$$P_{Q+}(n) \to \frac{1-(1-\rho^{-n})(a_Q/a)}{1-(a_Q/a)} \ \textit{and} \ P_{ab}(n) \to \frac{\epsilon}{1-(a_Q/a)}.$$

   (b) *(Square-root staffing rule) When $R_Q = R$ (thus, $\epsilon + \tau = 0$), we take the limit of large R while fixing the square-root coefficient c that satisfies $s = R + c\sqrt{R}$ and obtain*

$$P_{Q+}(n) \to \phi(c) \ \textit{and} \ P_{ab}(n) \to \epsilon\phi(c).$$

Proof of Proposition 2. Following the linear staffing representation, it is easy to analyze two extreme cases: $s < R_Q$ $(\leq R)$ (ED regime) and $s > R$ $(\geq R_Q)$ (QD regime). Using Lemmas 2 and 5, we obtain $P_W(n) \to 0$, $P_Q(n) \to 1$, $P_{Q+} = P_W(n) + P_Q(n) \to 1$, $P_{ab}(n) \to p$ for the ED regime; and $P_W(n) \to 0$, $P_Q(n) \to 0$, $P_{Q+} = P_W(n) + P_Q(n) \to 0$, $P_{ab}(n) \to 0$ for the QD regime.

We next consider the QED regime that exists in between the two extreme (ED and QD) regimes. If $R_Q \neq R$, the linear staffing rule following $R_Q < s = R + aR < R$ can achieve the QED regime. The properties of this QED regime are obtained using Lemmas 2 and 5: $P_W(n) \to -\frac{(1-\rho^{-n})(a_Q/a)}{1-(a_Q/a)}$, $P_Q(n) \to \frac{1}{1-(a_Q/a)}$, $P_{Q+} = P_W(n) + P_Q(n) \to \frac{1-(1-\rho^{-n})(a_Q/a)}{1-(a_Q/a)}$, and $P_{ab}(n) \to \frac{\varepsilon}{1-(a_Q/a)}$. (We omit the calculation since it is straightforward, although cumbersome.)

If $R_Q = R$ (i.e., $\epsilon + \tau = 0$), two extreme regimes are adjacent in the linear staffing representation. Thus, we utilize the finer square-root staffing representation to describe the properties of the QED regime that exists at the boundary of the ED and QD regimes. Using Lemmas 2 and 6, we obtain $P_W(n) \to 0$, $P_Q(n) \to \phi(c)$, $P_{Q+}(n) = P_W(n) + P_Q(n) \to \phi(c)$, and $P_{ab}(n) \to \epsilon\phi(c)$. □

Proposition 2 shows that there is no trade-off between the delay probability ($P_{Q+}(n)$) and the abandonment probability ($P_{ab}(n)$) in the asymptotic limit of R. In the two extreme regimes, ED and QD, both $P_{Q+}(n)$ and $P_{ab}(n)$ do not depend on n, implying that the number of seats n in the buffer space does not impact the performance indicators. For the QED regime that exists in between the two extreme regimes, we consider two scenarios:

(1) if $R \neq R_Q$, the linear staffing rule applies and tells us that as n increases, $P_{Q+}(n)$ could increase (since $a_Q > 0$, $a < 0$, and $\rho > 1$); and (2) if $R = R_Q$, the square-root staffing rule applies and tells us that $P_{Q+}(n)$ is not affected by n. $P_{ab}(n)$ remains the same as n increases for both scenarios. We conclude that providing a buffer space would not be beneficial in the asymptotic limit, which is in contrast to the non asymptotic limit case (Proposition 1).

5. Conclusions

We propose a queueing model with customer abandonment (reneging/balking) that incorporates a buffer space. We call this model the deferred abandonment model. We derive asymptotic formulae for performance indicators: the delay probability and the abandonment probability. We find that the provision of the buffer space may be worthwhile for smaller systems, but is not beneficial and in fact could be harmful for larger systems. This is because the buffer space may only exacerbate the delay probability without improving the throughput of the facility in an asymptotic limit.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The author is grateful for the helpful comments and suggestions provided by the review team, which greatly improved the presentation of this paper. The author appreciates Robert Hampshire and Alan Scheller-Wolf for their valuable feedback and support. The author acknowledges Teresa Rainone for her advice on one of the examples used in this paper. The author is also thankful to Mary Glavan and Jasim Ahmed for their helpful revisions of the paper. This paper is dedicated to the memory of Matthew S. for his genuine friendship.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Palm, R.C.A. *Research on Telephone Traffic Carried by Full Availability Groups*; Tele: Stockholm, Sweden, 1957; Volume 1.
2. Udagawa, K.; Nakamura, G. On a Queue in which Joining Customers Give up Their Services Halfway. *J. Ops. Res. Jpn.* **1957**, *1*, 59–76.
3. Haight, F.A. Queueing with reneging. *Metrika* **1959**, *2*, 186–197. [CrossRef]
4. Ancker, C.; Gafarian, A. Some queuing problems with balking and reneging. I. *Oper. Res.* **1963**, *11*, 88–100. [CrossRef]
5. Dong, J.; Feldman, P.; Yom-Tov, G.B. Service systems with slowdowns: Potential failures and proposed solutions. *Oper. Res.* **2015**, *63*, 305–324. [CrossRef]
6. Azriel, D.; Feigin, P.D.; Mandelbaum, A. Erlang-S: A data-based model of servers in queueing networks. *Manag. Sci.* **2019**, *65*, 4607–4635. [CrossRef]
7. Wu, C.; Bassamboo, A.; Perry, O. Service system with dependent service and patience times. *Manag. Sci.* **2019**, *65*, 1151–1172. [CrossRef]
8. Garnett, O.; Mandelbaum, A.; Reiman, M. Designing a call center with impatient customers. *Manuf. Serv. Oper. Manag.* **2002**, *4*, 208–227. [CrossRef]
9. Mandelbaum, A.; Zeltyn, S. *The Palm/Erlang—A Queue, with Applications to Call Centers*; Faculty of Industrial Engineering & Management, Technion: Haifa, Israel, 2005.
10. Sasanuma, K.; Scheller-Wolf, A. Approximate performance measures for a single station two-stage reneging queue. *Oper. Res. Lett.* **2021**, *49*, 212–217. [CrossRef]
11. Kelly, F.P. *Reversibility and Stochastic Networks*; John Wiley & Sons: New York, NY, USA, 1979.
12. Sasanuma, K.; Hampshire, R.; Scheller-Wolf, A. Markov chain decomposition based on total expectation theorem. *arXiv* **2019**, arXiv:1901.06780 2019.