ANZIAM J. **68, e17** (2026), 1–14
doi:10.1017/S1446181126100406

TRANSIENT ANALYSIS OF THE $M/M/K/\ell + M$ QUEUE

HRITIKA GUPTA [1,2]

(Received 7 December, 2025 UTC; accepted 11 May, 2026 UTC)

Abstract

The transient behaviour of an $M/M/k$ queue with finite capacity ℓ and impatient customers is analysed in this paper. Expressions for the abandonment probabilities and the waiting time distribution of a tagged caller are derived using Laplace transforms. Laplace transforms are then applied to the Kolmogorov forward equations, combined with numerical inversion, to compute the transition probabilities. By combining these results, the probability that a customer arriving at time t is served within a specified time window, as well as their probability of abandonment, can be determined. Numerical examples are provided throughout the paper to illustrate the results.

2020 *Mathematics subject classification*: 60K25.

Keywords and phrases: probability, abandonments, waiting time distributions, queueing, transient.

1. Introduction

Queueing systems are fundamental models for service operations, telecommunications and call centres. While stationary analysis provides valuable long-term insights, many practical systems spend a significant portion of their time in nonequilibrium states, making transient analysis essential for operational decision-making.

There has been substantial research on queueing systems under stationary conditions, but relatively fewer studies have focused on their transient behaviour.

Earlier work includes the work of Clarke [9], which provided an integral solution for the transient queue length distribution of the $M/M/1$ queue, and later Saaty [18]

[1]School of Mathematics and Statistics, The University of Melbourne, Parkville, 3010 Victoria, Australia
[2]Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Carlton, 3053 Victoria, Australia; e-mail: hgupta@uow.edu.au
© The Author(s), 2026. Published by Cambridge University Press on behalf of The Australian Mathematical Publishing Association Inc. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

obtained the Laplace transform of the transient transition probabilities for the $M/M/c$ queue. This was followed by contributions such as the Laplace–Stieltjes transform (LST) of the transient transition probabilities for bulk queues by Neuts [16] and an algorithm for computing transition probabilities by Grassmann [12]. Abate et al. [1] developed an algorithm to numerically invert multidimensional transforms, and applied it to find the transient queue length and workload distribution of the $M/G/1$ queue. Analytical expressions for the time-dependent distribution of $M/D/1/N$ queues initialised in an arbitrary deterministic state were derived by Garcia et al. [11].

A range of further studies have focused on transient state probabilities. These include transient probabilities for the $M/M/c$ queue with balking and reneging [3], time-dependent system-size probabilities for the $M/M/1$ queue under various conditions [4, 21], eigenvalue-based methods for $M/M/c$ queues with finite buffers [13] and matrix techniques for transient state probabilities of $M/M/1/N$ queues [5]. Performance measures using the stationary backlog carryover approach in finite-buffer queues with general inter-arrival and service distributions were provided by Stolletz and Lagershausen [20], while Knessl and van Leeuwaarden [15] used Green's functions and contour integrals to obtain transient distributions and first passage times for the $M/M/c+M$ queue. The time-dependent transition function for an $M/M/c$ priority queue with class-dependent service rates was derived by Selen and Fralix [19].

Most work on transient waiting time distributions has focused on single-server queues. These include the standard deviation and covariance function of transient waiting time distributions for the $GI/G/1$ queue [7], an integral solution for the $G/M/1$ priority queue [10] and closed-form Laplace transforms for the $G/G/1$ queue [6]. Lower bounds for the tails of steady-state and transient waiting time distributions in the $M/GI/s$ queue were established by Whitt [23], and Van Houdt and Blondia [22] proposed a steady-state based approximation method for the queue length and waiting time distributions in discrete-time systems.

To the best of our knowledge, no study has analysed the transient waiting time distribution and abandonment probabilities in an $M/M/k/\ell+M$ queue (an $M/M/k$ queue with finite capacity and impatient customers). The paper fills this gap by deriving explicit expressions for the transient waiting time distribution and abandonment probabilities for a tagged customer in an $M/M/k/\ell+M$ queue.

The paper is structured as follows. The model is described in Section 2. The waiting time distribution of a tagged customer in an $M/M/k/\ell+M$ queue is derived in Section 3 using the LST. In Section 4, the abandonment probability of the tagged caller is derived using a modified version of the LST expression obtained in the previous section, as well as by direct computation. It is shown that the probability that a customer in position i abandons before being served is identical to the probability that the first event after their arrival to the call centre is an abandonment.

A set of linear equations satisfied by the Laplace transform of the transient transition probabilities is derived in Section 5. This results in a second-order nonhomogeneous recursion, which can be solved using a tridiagonal solver and then numerically inverted using the Euler algorithm to obtain the transition probabilities. These measures can be

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

combined to calculate the probability that a customer arriving at time t abandons or is served within a specified time period, given the initial state of the system. Numerical illustrations that show how these probabilities evolve over time depending on the initial state are also presented.

2. The model

An $M/M/k/\ell + M$ queue is considered, that is, a queue with Poisson arrivals, exponential service time, exponential abandonment time, k servers and capacity ℓ so that no more than ℓ customers can be in the system at any time, with $\ell \geq k$. A customer is blocked if they call when all lines are busy. It is also assumed that a caller cannot abandon once the service has commenced. Customers are served on a first-come, first-served basis in a work-conserving manner.

Let λ be the arrival rate of customers, μ be the service rate and θ be the abandonment rate. The system is modelled as a continuous-time Markov chain [14]. Let $Y(t)$ denote the total number of customers in the system at time t for $t > 0$ with state space $S_y = \{0, 1, \ldots, \ell\}$.

From state j, transitions occur as follows. An arrival increases the state to $j + 1$ at rate λ when $j < \ell$; otherwise, arrivals are blocked. Departures occur either due to service completions or abandonment. If $j \leq k$, there are j customers in service and 0 customers in the queue. If $j > k$, there are k customers in service and $j - k$ customers in the queue. Hence, the number of customers in service is $\min(j, k)$ and the number of customers in the queue is $\max(0, j - k)$. This implies that the total rate from state j to $j - 1$ is given by

$$\min(j, k)\mu + \max(0, j - k)\theta.$$

Thus, the transition rates are

$$q_{j,j+1} = \lambda \mathbb{1}_{\{j<\ell\}}, \quad q_{j,j-1} = \min(j, k)\mu + \max(0, j - k)\theta.$$

3. The waiting time distribution of a tagged caller

The aim of this paper is to compute the waiting time distribution of arriving customers. The analysis begins by studying the waiting time of a tagged caller whose position in the system is known upon arrival.

Specifically, the focus is on the time a caller spends waiting in the queue before being assigned to an available server. Thus, the waiting time ends when the caller enters service. Attention is restricted to cases where the waiting time is strictly positive, that is, when the tagged caller joins the queue. If an arriving customer finds fewer than k customers in the system, they enter service immediately and therefore experience zero waiting time.

To analyse the waiting time, a second continuous-time Markov chain $X(t)$ is introduced to represent the position of the tagged caller in the system at time t, conditional on the caller being in the queue. The state space of $X(t)$, denoted by S_x,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

is given by the waiting positions $\{k+1, k+2, \ldots, \ell\}$. Since the position of the tagged caller can only decrease over time, $X(t)$ evolves as a pure death process.

The process $X(t)$ can be absorbed in two ways. The first occurs when the tagged caller enters service and then the absorption time is the waiting time of interest. The second occurs when the tagged caller abandons the queue, in which case the waiting time is taken to be infinite. Therefore, the waiting time is a defective random variable, as it can take the value ∞ with positive probability.

When $X(t) = i$, there are $i-1$ customers ahead of the tagged caller, of whom k are in service and $i-1-k$ are waiting in the queue. Service completions occur at rate $k\mu$, while customers ahead of the tagged caller abandon the queue at rate $(i-1-k)\theta$. Hence, the total rate of transition from state i to $i-1$ is given by $k\mu + (i-1-k)\theta$.

In addition, the tagged caller may abandon the queue at rate θ, in which case the process is absorbed. Therefore, the total transition rate out of state i is given by $k\mu + (i-k)\theta$.

Let τ_i denote the sojourn time in state i. Then, τ_i is exponentially distributed with rate $k\mu + (i-k)\theta$.

Let W_i denote the waiting time of a tagged caller who arrives in position i. If the caller abandons the system before entering service, then $W_i = \infty$. For $y > 0$, let $\Xi_i(y)$ denote the waiting time distribution of W_i, that is,

$$\Xi_i(y) = \mathbb{P}(W_i \le y).$$

Since W_i is a defective random variable, the limit of $\Xi_i(y)$ as $y \to \infty$ is equal to the probability that the caller is eventually served. The Laplace–Stieltjes transform (LST) of $\Xi_i(y)$, denoted by $\bar{\Xi}_i(s)$, is given by

$$\bar{\Xi}_i(s) = \int_0^\infty e^{-st}\, d\Xi_i(t) = \mathbb{E}(e^{-sW_i}).$$

The aim is to compute $\mathbb{E}(e^{-sW_i})$. If $i \le k$, that is, the tagged caller immediately enters service upon arrival, $W_i = 0$. Hence,

$$\mathbb{E}(e^{-sW_i}) = 1. \tag{3.1}$$

To compute $\mathbb{E}(e^{-sW_i})$ when the tagged caller joins the queue, a first-step analysis is used. When the tagged caller is in position i, at the next event, there are two possibilities.

- The tagged caller abandons with probability

$$\frac{\theta}{(i-k)\theta + k\mu},$$

  in which case $W_i = \infty$ and, hence, $e^{-sW_i} = 0$.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

- The position of the tagged caller decreases to $i-1$ with probability

$$\frac{(i-1-k)\theta + k\mu}{(i-k)\theta + k\mu},$$

due to either a service completion or abandonment of a customer ahead. In this case, $W_i = \tau_i + W_{i-1}$.

Conditioning on τ_i and the next event yield

$$\mathbb{E}(e^{-sW_i}) = \mathbb{E}(\mathbb{E}(e^{-sW_i} \mid \tau_i, X(\tau_i))).$$

Only the transition to $i-1$ contributes, since the abandonment case yields zero. Denote the probability density function of τ_i by

$$f_{\tau_i}(t) = ((i-k)\theta + k\mu)e^{-((i-k)\theta+k\mu)t}, \quad t \geq 0.$$

Thus,

$$\begin{aligned}\mathbb{E}(e^{-sW_i}) &= \int_0^\infty f_{\tau_i}(t)\mathbb{P}(X(\tau_i) = i-1)\mathbb{E}(e^{-s(W_{i-1}+t)} \mid \tau_i = t, X(\tau_i) = i-1)\,dt \\ &= \mathbb{P}(X(\tau_i) = i-1)\mathbb{E}(e^{-sW_{i-1}})((i-k)\theta + k\mu)\int_0^\infty e^{-((i-k)\theta+k\mu+s)t}\,dt \\ &= \frac{(i-1-k)\theta + k\mu}{(i-k)\theta + k\mu}\mathbb{E}(e^{-sW_{i-1}})\frac{(i-k)\theta + k\mu}{(i-k)\theta + k\mu + s}.\end{aligned}$$

Hence,

$$\mathbb{E}(e^{-sW_i}) = \frac{(i-1-k)\theta + k\mu}{(i-k)\theta + k\mu + s}\mathbb{E}(e^{-sW_{i-1}}).$$

Iterating the recursion and using (3.1) yields

$$\mathbb{E}(e^{-sW_i}) = \prod_{j=k+1}^{i}\frac{(j-1-k)\theta + k\mu}{(j-k)\theta + k\mu + s}. \tag{3.2}$$

It is known that the inverse LST of a function of the form $\alpha_1/(\beta_1 + s)$, where s is the Laplace variable, and α_1 and β_1 are constants such that $\alpha_1 \leq \beta_1$, is given by

$$F(t) = \frac{\alpha_1}{\beta_1}(1 - e^{-\beta_1 t}).$$

This is a cumulative distribution function of the exponential distribution with rate β_1 scaled by α_1.

To see this, note

$$\frac{\alpha_1}{\beta_1 + s} = \alpha_1\int_0^\infty e^{-st}e^{-\beta_1 t}\,dt$$

implying $dF(t) = \alpha_1 e^{-\beta_1 t}\,dt$ and, hence, the result.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

Now, the inverse LST of the more generalised formula $\prod_{j=1}^{i}\alpha_j/(\beta_j+s)$ is discussed, which mimics the final form of (3.2).

Each term in this product $\alpha_j/(\beta_j+s)$ can be written as $(\alpha_j/\beta_j)(\beta_j/(\beta_j+s))$, where $\beta_j/(\beta_j+s)$ is the LST of an exponential distribution with rate β_j.

The inverse LST of $\prod_{j=1}^{i}\beta_j/(\beta_j+s)$ is the distribution function of the hypoexponential distribution, that is, the distribution of the sum of independent exponential random variables with distinct rates $\beta_1,\beta_2,\ldots,\beta_i$, implying

$$F(t)=1-\sum_{j=1}^{i}\left(\left(\prod_{\substack{k=1\\k\neq j}}^{i}\frac{\beta_k}{\beta_k-\beta_j}\right)e^{-\beta_j t}\right)$$

[24]. The rates are also distinct for the final expression in (3.2).

Hence, the inverse LST of $\prod_{j=1}^{i}\alpha_j/(\beta_j+s)$ is given by

$$F(t)=\left(\prod_{j=1}^{i}\frac{\alpha_j}{\beta_j}\right)\left(1-\sum_{j=1}^{i}\left(\left(\prod_{\substack{k=1\\k\neq j}}^{i}\frac{\beta_k}{\beta_k-\beta_j}\right)e^{-\beta_j t}\right)\right). \tag{3.3}$$

It follows that $\Xi_i(y)$ has the same form as (3.3), with $\alpha_j=(j-1-k)\theta+k\mu$ and $\beta_j=(j-k)\theta+k\mu$. This expression allows us to calculate the time-dependent waiting time distribution of the customer with position i in the system. A numerical example and an accompanying plot are presented at the end of this section.

3.1. Other interpretations The quantity in (3.2),

$$\mathbb{E}(e^{-sW_i})=\prod_{j=k+1}^{i}\frac{(j-1-k)\theta+k\mu}{(j-k)\theta+k\mu+s},$$

has two useful interpretations.

- It can be viewed as a discounted reward for completing service. Suppose a customer receives a unit reward upon being served, with no other costs involved. Then, $\mathbb{E}(e^{-sW_i})$ represents the expected discounted reward earned by a tagged customer who arrives in position i in the system, where discounting occurs at rate s.
- It is also equal to the probability that the caller is served before an independent "killing" random variable $\mathcal{K}$ terminates the process, where $\mathcal{K}$ is an exponentially distributed random variable with rate s. This interpretation holds because the expression for the LST of the waiting time random variable is identical to that of $\mathbb{E}(\mathbb{1}_{\{W_i<\mathcal{K}\}})$. This interpretation is used in Section 4.

EXAMPLE 3.1. For the parameter values $\lambda=1,\mu=1/3,\theta=1/4,k=5$ and $\ell=20$, Table 1 presents the values of $\Xi_i(0.5)$, that is, the probability of serving a customer in 0.5 time units given that they arrive in position i in the system.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

TABLE 1. $\Xi_i(0.5)$ for $i \in \{1, 2, \ldots, 20\}$, $\lambda = 1, \mu = 1/3, \theta = 1/4, k = 5$ and $\ell = 20$.

i	$\Xi_i(0.5)$	i	$\Xi_i(0.5)$
1	1.000	11	9.65×10^{-4}
2	1.000	12	2.02×10^{-4}
3	1.000	13	3.998×10^{-5}
4	1.000	14	7.58×10^{-6}
5	1.000	15	1.38×10^{-6}
6	0.537	16	2.45×10^{-7}
7	0.208	17	4.21×10^{-8}
8	0.066	18	7.07×10^{-9}
9	0.018	19	1.16×10^{-9}
10	0.004	20	1.88×10^{-10}

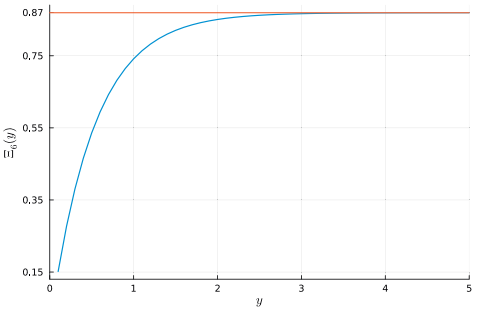


FIGURE 1. Plot of $\Xi_6(y)$, the probability that a caller in position 6 in the system is served within y time units.

As expected, the probability of being served within 0.5 time units is 1 when the tagged caller is in position less than or equal to k, it drops to 0.537 when the tagged caller is in position $k + 1$ and decreases further as the position of the tagged caller in the system increases.

For the same set of parameter values, Figure 1 shows the plot of $\Xi_6(y)$ against y. The function $\Xi_6(y)$ gives the probability that a caller who arrives in position 6 in the system is served within y time units. Its limiting value, which represents the probability that they are eventually served, is 0.87. This implies an abandonment probability of 0.13.

4. The abandonment probability of a tagged caller

Using the second interpretation from Section 3.1, the abandonment probability for the tagged caller can be derived. For now, assume that the customers ahead of the tagged caller may abandon, but the tagged caller does not abandon when defining its

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

waiting time. Let $\mathcal{W}_i$ denote the waiting time of a customer initially in position i under this modified system.

Under this assumption, the expression for the LST in (3.2) becomes

$$\mathbb{E}(e^{-s\mathcal{W}_i}) = \prod_{j=k+1}^{i} \frac{(j-1-k)\theta + k\mu}{(j-1-k)\theta + k\mu + s}. \tag{4.1}$$

The change in the denominator reflects the fact that the tagged caller is not exposed to abandonment when computing this transform.

Now, introduce an independent exponential random variable $\mathcal{K}$ with rate s. Then,

$$\mathbb{P}(\mathcal{W}_i < \mathcal{K}) = \mathbb{E}(e^{-s\mathcal{W}_i}),$$

since $\mathcal{K}$ is independent of $\mathcal{W}_i$.

Setting $s = \theta$ gives the tagged customer an independent exponential patience clock with the original system abandonment rate. Substituting $s = \theta$ in (4.1) yields a telescoping product that simplifies to

$$\frac{k\mu}{(i-k)\theta + k\mu},$$

which is the probability that a customer who enters in position i is eventually served.

The probability that a tagged customer abandons the system, given that they arrive in position i, can also be computed directly. Let ζ_i denote this probability.

For positions where the customer is served immediately,

$$\zeta_i = 0. \tag{4.2}$$

For positions with strictly positive waiting time, a first-step analysis is again used. At the first step, two outcomes are possible—either the tagged customer abandons immediately or they move one position closer to service. This gives the recursion

$$\zeta_i = \frac{\theta}{(i-k)\theta + k\mu} + \frac{(i-1-k)\theta + k\mu}{(i-k)\theta + k\mu}\zeta_{i-1}.$$

This can be rewritten as

$$1 - \zeta_i = \frac{(i-1-k)\theta + k\mu}{(i-k)\theta + k\mu}(1 - \zeta_{i-1}).$$

Iterating this recursion from i down to k and using $1 - \zeta_k = 1$ from (4.2) yields the telescoping product

$$1 - \zeta_i = \prod_{j=k+1}^{i} \frac{(j-1-k)\theta + k\mu}{(j-k)\theta + k\mu} = \frac{k\mu}{(i-k)\theta + k\mu},$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

TABLE 2. ζ_i for $i \in \{1, 2, \ldots, 20\}$ and the parameter values $\lambda = 1, \mu = 1/3, \theta = 1/4, k = 5$ and $\ell = 20$.

i	ζ_i	i	ζ_i
1	0.000	11	0.474
2	0.000	12	0.512
3	0.000	13	0.545
4	0.000	14	0.574
5	0.000	15	0.600
6	0.130	16	0.623
7	0.231	17	0.643
8	0.310	18	0.661
9	0.375	19	0.677
10	0.429	20	0.692

which is the same as (4.1). Hence, the probability that a tagged customer in position $i > k$ eventually abandons is

$$\zeta_i = \frac{(i-k)\theta}{(i-k)\theta + k\mu}. \tag{4.3}$$

The expressions in (4.3) and (4.1) are of interest because they show that the probability that a customer in position i abandons before being served is identical to the probability that the first event after they enter the queue is an abandonment.

EXAMPLE 4.1. The values for ζ_i for $i \in \{1, 2, \ldots, 20\}$ are presented in Table 2 for the parameter values $\lambda = 1$, $\mu = 1/3$, $\theta = 1/4$, $k = 5$ and $\ell = 20$. The values imply that when the tagged caller is sixth in the system, the probability of abandonment is 0.130 (consistent with the value obtained from Figure 1). If the tagged caller is seventh in the system, this probability increases to 0.231 and it continues to rise as the position moves further back.

5. Time-dependent transition probabilities

A method for computing transient transition probabilities is illustrated in this section. Together with the results in Sections 3–4, this allows us to answer questions such as, given the initial state of the system, what is the probability that a customer arriving at time t will abandon or be served within a specified time period.

For this, the original Markov chain $Y(t)$ defined in Section 2 is considered.

If $P^{(t)} = [p_{i_0,j}(t)]_{i_0,j \in S_y}$ is the probability matrix for the Markov chain $Y(t)$ and Q is the generator matrix, then the aim is to find an expression for $P^{(t)} = \exp(tQ)$ [8].

Let $p_{i_0,j}(t) = 0$ whenever $i_0 \notin S_y$ or $j \notin S_y$. Let the total rate of the system at state i_0 be denoted by $\kappa(i_0)$, then $\kappa(i_0) = \lambda \mathbb{1}_{i_0 < \ell} + (\min(i_0, k)\mu + \max(0, i_0 - k)\theta)\mathbb{1}_{i_0 > 0}$. For $i_0, j \in S_y$, the forward Kolmogorov equations are

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

$$\frac{dp_{i_0,j}(t)}{dt} = \lambda p_{i_0,j-1}(t) + (\max(0, j+1-k)\theta + \min(j+1, k)\mu)p_{i_0,j+1}(t) - \kappa(j)p_{i_0,j}(t).$$

Taking Laplace transforms on both sides yields

$$s\tilde{p}_{i_0,j}(s) - p_{i_0,j}(0) = \lambda\tilde{p}_{i_0,j-1}(s) + (\max(0, j+1-k)\theta + \min(j+1, k)\mu)\tilde{p}_{i_0,j+1}(s) - \kappa(j)\tilde{p}_{i_0,j}(s)$$

and $p_{i_0,j}(0) = \mathbb{1}_{i_0=j}$. Rearranging gives

$$\tilde{p}_{i_0,j}(s) = \frac{\lambda\tilde{p}_{i_0,j-1}(s) + (\max(0, j+1-k)\theta + \min(j+1, k)\mu)\tilde{p}_{i_0,j+1}(s) + \mathbb{1}_{i_0=j}}{\kappa(j) + s}.$$

The system can be written in matrix form as

$$\boldsymbol{A}\tilde{\boldsymbol{p}}_{i_0} = \boldsymbol{b}_{i_0},$$

where

$$\tilde{\boldsymbol{p}}_{i_0} = \begin{pmatrix} \tilde{p}_{i_0,0}(s) \\ \tilde{p}_{i_0,1}(s) \\ \tilde{p}_{i_0,2}(s) \\ \vdots \\ \tilde{p}_{i_0,\ell}(s) \end{pmatrix}, \qquad \boldsymbol{b}_{i_0} = \begin{pmatrix} \mathbb{1}_{i_0=0} \\ \mathbb{1}_{i_0=1} \\ \mathbb{1}_{i_0=2} \\ \vdots \\ \mathbb{1}_{i_0=\ell} \end{pmatrix}.$$

The tridiagonal matrix $\boldsymbol{A}$ has entries

$$A_{j,j} = \kappa(j) + s \quad \text{and} \quad A_{j,j-1} = -\lambda$$

for $j \geq 1$, and

$$A_{j,j+1} = -(\max(0, j+1-k)\theta + \min(j+1, k)\mu)$$

for $j \leq \ell - 1$, with all other entries equal to zero. Explicitly, it can be displayed as

$$\boldsymbol{A} = \begin{pmatrix} \kappa(0)+s & -(\theta\cdot\max(0,1-k)+\mu\cdot\min(1,k)) & \cdots & 0 \\ -\lambda & \kappa(1)+s & \cdots & 0 \\ 0 & -\lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \kappa(\ell)+s \end{pmatrix}.$$

Since $\boldsymbol{A}$ is tridiagonal, the linear system can be solved using a tridiagonal solver based on the Thomas algorithm [17]. It provides a computational improvement compared with general matrix inversion, particularly for large values of ℓ.

Then, $\tilde{p}_{i_0,j}(s)$ can be inverted using the Euler method [2] to obtain the transition probabilities.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

TABLE 3. $p_{3,j}(60)$ for the parameter values $\lambda = 1, \mu = 1/3, \theta = 1/4, k = 5$ and $\ell = 20$.

j	$p_{3,j}(60)$	j	$p_{3,j}(60)$
0	0.049	11	4.04×10^{-4}
1	0.149	12	1.18×10^{-4}
2	0.222	13	3.23×10^{-5}
3	0.222	14	8.24×10^{-6}
4	0.167	15	1.98×10^{-6}
5	0.100	16	4.48×10^{-7}
6	0.052	17	9.60×10^{-8}
7	0.024	18	1.95×10^{-8}
8	0.010	19	3.78×10^{-9}
9	0.004	20	6.97×10^{-10}
10	0.001		

In the Euler inversion algorithm, the Laplace transform must be evaluated at complex values of the Laplace parameter of the form

$$s_r = \frac{A + 2r\pi\iota}{2t}, \quad r = 0, 1, \ldots, n + m + 1.$$

Accordingly, the coefficients of the tridiagonal system become complex-valued, but the Thomas algorithm applies without modification since it only involves addition, multiplication and division operations, which are well defined for complex numbers. All computations are performed using complex double-precision arithmetic.

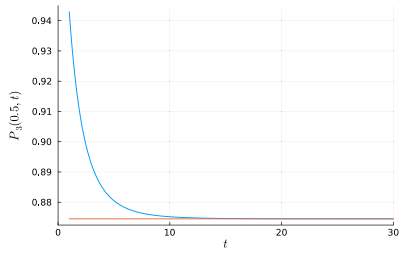
The parameters used in the Euler inversion follow the recommendations of [2]. Specifically, $A = 18.4, n = 15$ and $m = 11$.

The truncation level n determines the number of terms used in the initial alternating series, while m specifies the number of additional terms used in the Euler step. The error is estimated using the difference between successive Euler approximations, $|E(m, n + 1, t) - E(m, n, t)|$, and this quantity is used as an indicator of truncation accuracy.

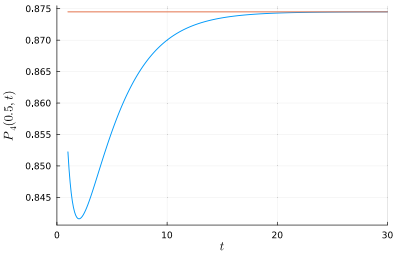
For Example 5.1, the resulting error estimates decrease rapidly and are of the order of 10^{-10} in the early iterations and eventually reach magnitudes around 10^{-17}. This indicates that the numerical inversion is accurate and that the truncation levels $n = 15$ and $m = 11$ are sufficient for the parameter values considered.

EXAMPLE 5.1. Table 3 presents the values of $p_{3,j}(60)$ for the parameter values $\lambda = 1, \mu = 1/3, \theta = 1/4, k = 5$ and $\ell = 20$, that is, the probability mass function for the state of the call centre at $t = 60$, given the call centre has 3 customers at time 0. Note that the probabilities sum to approximately one (up to numerical rounding), confirming the consistency of the numerical inversion procedure.

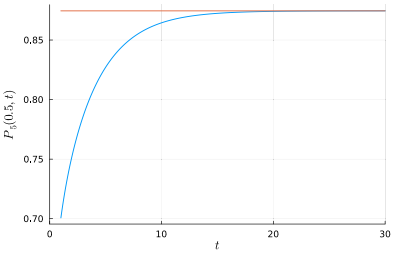
Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406



(a) $i_0 = 3$



(b) $i_0 = 4$



(c) $i_0 = 5$

FIGURE 2. Plot of $P_{i_0}(0.5, t)$ against t.

These values imply that if there are 3 customers in the system at the start, the probability of finding the call centre empty at $t = 60$ is 0.049, the probability of finding one customer in the system is 0.149 and so on.

To validate the Laplace inversion procedure [2], the transient state probabilities are also computed by calculating the matrix exponential of the generator matrix. For the finite-state system considered, the inversion-based results matched the direct computation with a maximum absolute error of order 10^{-9}.

Now, let $P_{i_0}(y, t)$ denote the probability that a caller arriving at time t is served within y time units, conditional on there being i_0 customers in the system at time 0. The system has finite capacity ℓ, so if an arrival occurs when the system is already in state ℓ, the customer is blocked and is not included in the performance measure.

The probability $P_{i_0}(y, t)$ can be computed using the results in Sections 3 and 5 as follows:

$$P_{i_0}(y, t) = \sum_{j=1}^{\ell} \Xi_j(y) p_{i_0, j-1}(t).$$

Here, $p_{i_0,j-1}(t)$ is the probability that the system is in state $j - 1$ at time t, given that the initial state at time 0 is i_0, and $\Xi_j(y)$ is the probability that a customer is served within y time units, given that they are in position j in the system upon arrival.

Figure 2 shows the plots of $P_{i_0}(y, t)$ against t for $y = 0.5$ and three different initial states $i_0 \in \{3, 4, 5\}$.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

Three distinct behaviours corresponding to the different initial values are observed. For $i_0 = 3$, the probability of service within 0.5 time units decreases with t. For $i_0 = 4$, the probability initially decreases for arrivals near time 0 and subsequently increases. For $i_0 = 5$, the probability increases over the entire range of t. In all cases, the curves converge to the same limiting value, corresponding to the stationary probability of being served within 0.5 time units. The abandonment probability for a caller arriving at time t can be obtained in the same way and displays a similar but opposite trend.

6. Conclusion

This paper analysed the transient behaviour of the $M/M/k/\ell + M$ queue with impatient customers. Explicit expressions were derived for the waiting time distribution and abandonment probability of a tagged caller using LST and first-step analysis.

Transient transition probabilities were computed by applying Laplace transforms to the Kolmogorov forward equations and numerically inverting the resulting transforms using the Euler algorithm. The transient state probabilities were combined with the waiting time results to evaluate quantities such as the probability that a customer arriving at time t is served within a specified time window.

The numerical examples illustrated how the transient performance of the system depends on the initial state and evolves towards the stationary regime over time.

Acknowledgements

This research was funded by the Melbourne Research Scholarship and the Australian Government through the Australian Research Council Industrial Transformation Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Project ID IC200100009. It was also supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative. The author is also funded by the OPTIMA Postgraduate Research Dissemination Award.

The author gratefully acknowledges Prof Peter Taylor and Dr Mark Fackrell for their guidance and support throughout this project.

References

[1] J. Abate, G. L. Choudhury and W. Whitt, "Calculation of the *GI/G*/1 waiting-time distribution and its cumulants from Pollaczek's formulas", *Arch. Elektron. Ubertr. (AEU)* **47** (1993) 311–321; https://api.semanticscholar.org/CorpusID:13917318.

[2] J. Abate and W. Whitt, "Numerical inversion of Laplace transforms of probability distributions", *ORSA J. Comput.* **7** (1995) 36–43; doi:10.1287/ijoc.7.1.36.

[3] R. O. Al-Seedy, A. A. El-Sherbiny, S. A. El-Shehawy and S. I. Ammar, "Transient solution of the $m/m/c$ queue with balking and reneging", *Comput. Math. Appl.* **57** (2009) 1280–1285; doi:10.1016/j.camwa.2009.01.017.

[4] S. I. Ammar, "Transient analysis of an *M*/*M*/1 queue with impatient behavior and multiple vacations", *Appl. Math. Comput.* **260** (2015) 97–105; doi:10.1016/j.amc.2015.03.066.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406

[5] S. I. Ammar, A. A. El-Sherbiny, S. A. El-Shehawy and R. O. Al-Seedy, "A matrix approach for the transient solution of an *M*/*M*/1/*N* queue with discouraged arrivals and reneging", *Int. J. Comput. Math.* **89** (2012) 482–491; doi:10.1080/00207160.2011.637553.

[6] D. J. Bertsimas and D. Nakazato, "Transient and busy period analysis of the *GI*/*G*/1 queue: the method of stages", *Queueing Syst. Theory Appl.* **10** (1992) 153–184; doi:10.1007/BF01159205.

[7] N. Blomqvist, "On the transient behaviour of the *GI*/*G*/1 waiting-times", *Scand. Actuar. J.* **1970**(3–4) (1970) 118–129; doi:10.1080/03461238.1970.10405658.

[8] K. Borovkov, *Elements of Stochastic Modelling* (World Scientific Publishing Company, Hackensack, NJ, 2014).

[9] A. B. Clarke, "A waiting line process of Markov type", *Ann. Math. Stat.* **27** (1956) 452–459; doi:10.1214/aoms/1177728268.

[10] G. Dalen and B. Natvig, "On the transient waiting times for a *GI*/*M*/1 priority queue", *J. Appl. Probab.* **17** (1980) 227–234; doi:10.2307/3212939.

[11] J.-M. Garcia, O. Brun and D. Gauchard, "Transient analytical solution of *M*/*D*/1/*N* queues", *J. Appl. Probab.* **39**(4) (2002) 853–864; doi:10.1239/jap/1037816024.

[12] W. Grassmann, "Transient solutions in Markovian queues: an algorithm for finding them and determining their waiting-time distributions", *Eur. J. Oper. Res.* **1** (1977) 396–402; doi:10.1016/0377-2217(77)90049-2.

[13] W. K. Grassmann and J. Tavakoli, "Transient solutions for multi-server queues with finite buffers", *Queueing Syst. Theory Appl.* **62** (2009) 35–49; doi:10.1007/s11134-009-9113-1.

[14] L. Kleinrock, *Queueing Systems, Volume I: Theory* (Wiley, New York, USA, 1975).

[15] C. Knessl and J. S. H. van Leeuwaarden, "Transient analysis of the Erlang A model", *Math. Methods Oper. Res.* **82** (2015) 143–173; doi:10.1007/s00186-015-0498-9.

[16] M. F. Neuts, "A general class of bulk queues with Poisson input", *Ann. Math. Stat.* **38** (1967) 759–770; doi:10.1214/aoms/1177698869.

[17] A. Quarteroni, R. Sacco and F. Saleri, *Numerical Mathematics*, Vol. **37** (Springer Science & Business Media, Berlin, 2006).

[18] T. L. Saaty, "Time-dependent solution of the many-server Poisson queue", *Oper. Res.* **8**(6) (1960) 755–772; doi:10.1287/opre.8.6.755.

[19] J. Selen and B. Fralix, "Time-dependent analysis of an *M/M/c* preemptive priority system with two priority classes", *Queueing Syst. Theory Appl.* **87** (2017) 379–415; doi:10.1007/s11134-017-9541-2.

[20] R. Stolletz and S. Lagershausen, "Time-dependent performance evaluation for loss-waiting queues with arbitrary distributions", *Int. J. Prod. Res.* **51** (2013) 1366–1378; doi:10.1080/00207543.2012.678946.

[21] R. Sudhesh and L. F. Raj, "Computational analysis of stationary and transient distribution of single server queue with working vacation", in: *International Conference on Computing and Communication Systems*, Volume 269 (eds P. V. Krishna, M. R. Babu and E. Ariwa) (Springer, Berlin, Heidelberg, 2011) 480–489; doi:10.1007/978-3-642-29219-4_55.

[22] B. Van Houdt and C. Blondia, "Approximated transient queue length and waiting time distributions via steady state analysis", *Stoch. Models* **21**(2–3) (2005) 725–744; doi:10.1081/STM-200056027.

[23] W. Whitt, "The impact of a heavy-tailed service-time distribution upon the M/GI/S waiting-time distribution", *Queueing Syst. Theory Appl.* **36** (2000) 71–87; doi:10.1023/A:1019143505968.

[24] G. P. Yanev, "Exponential and hypoexponential distributions: some characterizations", *Mathematics* **8**(12) (2020) 2207; doi:10.3390/math8122207.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:26:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1446181126100406