

Article

Transient Analysis of a Finite Queueing System with Bulk Arrivals in IoT-Based Edge Computing Systems

Shensheng Tang 

Department of Electrical and Computer Engineering, St. Cloud State University, St. Cloud, MN 56301, USA; stang@stcloudstate.edu

Abstract: Queueing models can be used for making decisions about the resources required to provide high quality service. In this paper, a finite capacity single server queueing model with bulk arrivals is studied in IoT-based edge computing systems. The transient analysis of the model is carried out and the transient analytical solution to the system is derived with a group of recursive coefficients by using the ordinary differential equations (ODEs) technique. From which the steady-state probabilities are solved. Then, some performance metrics of interest are derived along with numerical results. Although the paper is initiated from the IoT based edge computing platform, the proposed system modeling and analysis method can be extended to more general situations such as telecommunication, manufacturing, transportation, and many other areas that are closely related to people's daily lives.

Keywords: queueing model; bulk arrivals; transient analysis; steady state probabilities; ODE technique; performance metrics



Citation: Tang, S. Transient Analysis of a Finite Queueing System with Bulk Arrivals in IoT-Based Edge Computing Systems. *IoT* **2022**, *3*, 435–449. <https://doi.org/10.3390/iot3040023>

Academic Editor: Yann-Gaël Guéhéneuc

Received: 17 October 2022

Accepted: 15 November 2022

Published: 17 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, smart applications have emerged in the Internet of Things (IoT) based computing architectures. The basic principle of IoT is to connect various data creation or accumulation devices using technologies such as smart sensors, actuators, radio-frequency identification (RFID), and mobile devices, through which these devices can communicate with each other. With the rapid growth of IoT devices and smart applications, the management among the service level agreement, quality of service (QoS) guarantee, and computing cost becomes more and more difficult. Edge computing is promising approach that can be used to increase the efficiency of computing platforms. As a middle layer between the IoT devices and cloud levels, the edge layer with computing, storage, and networking capabilities can make users benefit from faster, more reliable services and organizations benefit from the flexibility of hybrid cloud computing.

Queueing models are widely used for the system modeling and performance evaluation. However, traditional queueing systems such as $M/M/1$, $M/M/c$, and $M/M/c/K$ [1] usually assume that the customers arrive singly at a service facility, which may not always reflect realistic situations. In reality, the customers or incoming traffic tasks may arrive to the system in groups. For example, in an IoT edge/cloud computing application for wildlife monitoring, the wild animals' behavior in situ, animal movement patterns, habitat utilization, and population demographics will be recorded and tracked by multiple types of sensors and sent to the edge server for processing when a predefined event is triggered. The IoT traffic generated by multiple types of sensors such as sound sensors, image sensors and video sensors will arrive at the edge server in bulk form due to a predefined event triggering mechanism. For another example, in the Internet of things (IoT) edge/cloud computing platform, one type of event-triggered IoT system generates the IoT traffic with bulk arrivals, where the IoT traffic generated by multiple sensors arrives at the edge server in groups when an event is triggered or the time has been scheduled beforehand.

Here list a few more examples as follows.

- Under the coronavirus pandemic environment, in a college library or bookstore, the number of entering students (either individual or student group) will be restricted after the number of students in the building is balanced, where the arrival students should wait in a queue in front of the door until a college staff allows a certain number of them to enter after the same number of students leave. In other words, when one student leaves, the head of line student in the queue will enter; when two students leave, the first two students in the queue will enter; and so on.
- In a doctor's clinic, the number of daily appointments and patient companions (i.e., visitors arriving at the same time) will be restricted due to the limited space.
- In transportation processes involving buses, airplanes, trains, ships, and elevators, where customers do not arrive singly, but in groups or bulk.

More examples can be listed: people going to a restaurant, visitors going to a Disneyland, letters arriving at a post office, etc. In all the above examples, we have observed a common phenomenon—the arrival of customers can be single or in groups and the group size may be a random variable or a fixed number. This arises a new class of queueing models—bulk queues (or batch queues) [2]. This is different from the ordinary queueing problems where it is often assumed that customers arrive singly [3,4].

The queueing models with bulk arrivals have been studied extensively in the literature [5–13]. In [5], the effect on a simple, single-server queueing system was investigated in which customers arrive in groups of a fixed number, but are served individually. In [6], a queueing system with infinite number of servers and batch arrivals was proposed where the joint behavior of the queue length, the number of customers who arrive during a certain period, and the total occupation time of the servers were studied. In [7], the behavior of a first-come-first-served queueing network with batch arrivals of variable size was studied and the Laplace transforms of the probability generating functions for the queue length were derived along with the steady-state results. In [8], a numerical method was developed for evaluating the distribution of the delay encountered by a customer in a time-inhomogeneous, single server queue with batch arrivals. In [9], a last-come, first-served queueing discipline and batch arrivals generated by a finite number of non-exponential sources was studied where a closed-form expression is derived for the steady-state queue length distribution.

In [10], an $M^k/M/\infty$ queue with k heterogeneous customers in a batch was proposed and the joint generating function of the number of customers of type i being served in the system in steady state was derived explicitly. In [11], a bulk arrival queueing model with fuzzy parameters and varying batch sizes was developed by a nonlinear programming approach to derive the membership functions of the steady-state performance measures. In [12], the authors studied the behavior of a batch arrival queueing system of a single server providing service in two modes with different service rates. The server may take a vacation or be subject to random breakdown. When the server faces breakdown, the customer in service will return back to the head of the queue and waits until the repair process is completed. In addition, the customers waiting for service may renege (leave the queue) when the server fails or takes vacation. In [13], the authors studied a single server bulk arrival queue system with batch size dependent service and working vacation, where the server provides service in two service modes depending upon the queue length. The server provides single service if the queue length is at least ' a ' while fixed batch service if the queue length is at least ' k ' ($k > a$). The probability generating function of the queue length was obtained by using supplementary variable technique.

Most of the above bulk arrival queueing systems considered infinite capacity, i.e., the number of customers allowed in the system is infinite. However, in practice many queueing systems have a constraint of capacity, i.e., there is a limit to the number of customers that may be in the queue or system [14,15]. An arriving customer who finds the system full cannot enter but leave the system immediately. In this case, there is a distinction between

the arrival rate (i.e., the number of arrivals per time unit) and the effective arrival rate (the number of arrivals that successfully enter the system per time unit).

On the other hand, most traditional queueing analysis studies the system behavior in steady state due to tractable analysis. However, in practice there are many classes of queueing systems in which a transient analysis is required [16]. As in many cases, the considered queueing system never reaches steady state; the steady-state simulation results do not accurately portray the system behavior. Even though a system can reach steady state, the steady-state results are obtained by running the system for long periods of time, which greatly nullifies the impact of initial conditions [17] and thus makes little known about the transient behavior.

There are quite a few studies in the literature on the transient analysis of queueing systems [18–22], which are summarized as follows. In [18], the difference equations that are satisfied by the Laplace transforms of the state probabilities at finite time were solved for an M/M/1/N queue and the state probabilities were thus obtained. In [19], a generating function approach along with the inversion of the generating function was used to obtain the transient probabilities of the M/M/1 queueing system. In [20], a simple series form was obtained for the transient state probabilities of a single server Markovian queue with finite waiting space, where the coefficients in the series satisfy iterative recurrence relations which enable fast and accurate numerical computations. In [21], an analytical expression of the time-dependent probability distribution of M/D/1/N queues initialized in an arbitrary deterministic state was derived and a simple analytical expression of the differential equation governing the transient average traffic which only involves probabilities of boundary states. In [22], a theoretical application of transient queueing analysis was provided for military air traffic control through the M/M/1 and the more general M/M/s queues. In [23], an analysis of the number of losses (caused by the buffer overflows) was presented in a finite-buffer queue with batch arrivals and autocorrelated inter-arrival times.

In this paper, a finite capacity single server queueing model with bulk arrivals is studied. The transient analysis of the model is carried out. By using the ordinary differential equations (ODEs) technique, we derive the transient analytical solution to a group of first-order nonhomogeneous linear ODEs of the queueing model with a group of recursive coefficients. From which the steady-state probabilities can be easily obtained when $t \rightarrow \infty$. We also develop some performance metrics of interest and perform the numerical evaluation for the metrics. The proposed system modeling and analysis method can provide more in-depth understanding for its applications to different fields including healthcare, telecommunication, commerce, manufacturing, transportation, and many other areas that are closely related to people's daily lives. Although the paper is initiated from the IoT based edge computing platform, clearly the involved modeling and analysis method can be applied to more general situations.

The remainder of the paper is organized as follows. Section 2 describes the general model of the system and assumptions. Section 3 analyzes the system model and derives the transient and steady state solutions along with a group of recursive formulas. Section 4 presents a case study for the system. Section 5 derives some system performance metrics of interest. Section 6 presents numerical results. Finally, the paper is concluded in Section 7.

2. Stochastic Queueing Model

The considered single server queueing model has bulk arrivals and finite waiting areas. The model can be mapped to many different application scenarios such as the aforementioned IoT based edge system and the event-triggered IoT system. The model under investigation is based on the following assumptions:

- (i) The single-server queueing system is finite with capacity of n . That is, the queue length or the maximum number of waiting places is $n - 1$.
- (ii) The customers arrive at a service facility in batches in accordance with a Poisson process with mean arrival rate λ .

- (iii) The number of arrivals may be either individuals or groups with random size, described by a random variable X with distribution given by $a_i = P(X = i)$, $i \geq 1$, where i is the number of customers in a group. If the group of customers arriving in the system finds j customers there, the whole group will enter the system when $i \leq n - j$; and leave the system when $i > n - j$.
- (iv) The service time of customers is a random variable with negative exponential distribution with parameter μ .
- (v) The queue discipline is first come first served (FCFS) by the arrivals and random inside the group.
- (vi) The arrivals, service times and batch sizes are mutually independent.

The notations of different arrival rates, time variables, and probabilities are listed in the nomenclature. The state of the system is determined by the number of customers in the system including the one in service. The state-transition diagram of the queueing model is shown in Figure 1. Let $p_k(t)$, $0 \leq k \leq n$, be the probability that k customers are present in the system at time t . Then, the transient system equations can be described by the following first-order nonhomogeneous linear ODEs:

$$p'_k(t) = -\left(\lambda \sum_{i=1}^n a_i\right)p_k(t) + \mu p_{k+1}(t), \quad k = 0 \tag{1}$$

$$p'_k(t) = -\left(\mu + \lambda \sum_{i=1}^{n-k} a_i\right)p_k(t) + \lambda \sum_{i=0}^{k-1} a_{k-i}p_i(t) + \mu p_{k+1}(t), \quad 1 \leq k \leq n - 1 \tag{2}$$

$$p'_k(t) = -\mu p_k(t) + \lambda \sum_{i=0}^{k-1} a_{k-i}p_i(t), \quad k = n \tag{3}$$

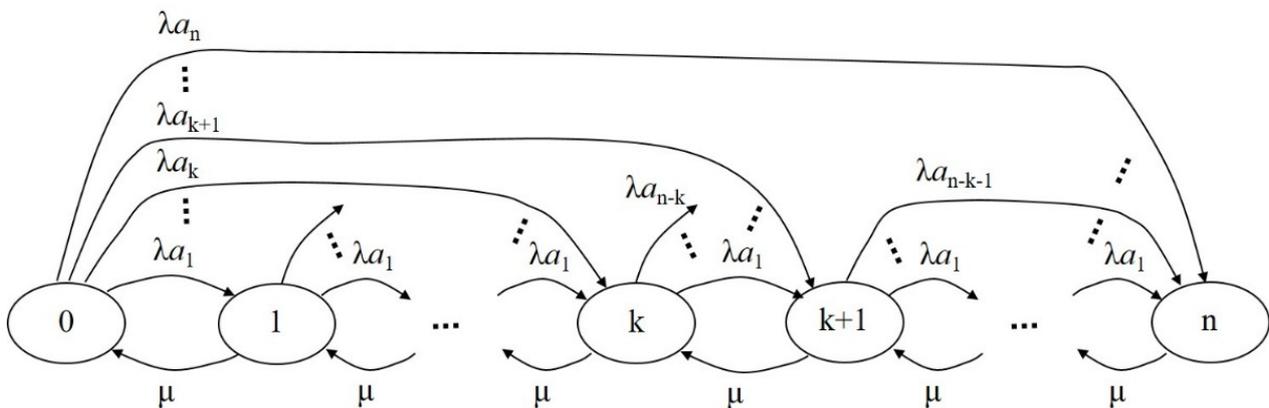


Figure 1. The single server queueing model with finite capacity and bulk arrivals.

When the batch size X is a uniform random variable, the probability of different batch size will be equally probable, i.e., $a_i = a$, $1 \leq i \leq n$. By applying the normalization condition

$$\sum_{k=0}^n p_k(t) = 1 \tag{4}$$

the above ODEs can be re-written as follows.

$$p'_0(t) = -\lambda n a p_0(t) + \mu p_1(t), \tag{5}$$

$$p'_k(t) = -[\mu + \lambda(n - k)a]p_k(t) + \lambda a \left[1 - \sum_{i=k}^n p_i(t)\right] + \mu p_{k+1}(t), \quad 1 \leq k \leq n - 1 \tag{6}$$

$$p'_n(t) = -\mu p_n(t) + \lambda a [1 - p_n(t)]. \tag{7}$$

It is observed that the above system, described by either ODEs (1)–(3) or ODEs (5)–(7), can be solved from the last ODE to the first with the obtained solutions applied to the solution of the next ODE. For the ease of presentation purpose, in the following we will focus the system solution on the ODEs (5)–(7).

3. Model Analysis and Solutions

In this section, we perform the analysis of the queueing model and derive the transient and steady state solutions to it. From the ODE (7), we have

$$p'_n(t) + (\mu + \lambda a)p_n(t) = \lambda a \tag{8}$$

By using the route ODE solving method [24], we obtain the solution of $p_n(t)$:

$$p_n(t) = C(n, 0) + C(n, 1)e^{-(\mu + \lambda a)t} \tag{9}$$

where the coefficient $C(n, 0) = \frac{\lambda a}{\mu + \lambda a}$, and $C(n, 1)$ is a real constant determined by the initial condition of the ODE.

Consider the next ODE with $k = n - 1$ and re-write the ODE as

$$p'_{n-1}(t) + (\mu + 2\lambda a)p_{n-1}(t) = (\mu - \lambda a)p_n(t) + \lambda a \tag{10}$$

Substituting (9) into (10) and solve it as:

$$p_{n-1}(t) = C(n - 1, 0) + C(n - 1, 1)e^{-(\mu + \lambda a)t} + C(n - 1, 2)e^{-(\mu + 2\lambda a)t} \tag{11}$$

where the coefficients $(n - 1, i)$, $i = 0, 1, 2$, are real constants determined by the initial condition of the ODE.

Similarly, the rest ODEs in (6) are solved with the form:

$$p_{n-k}(t) = C(n - k, 0) + \sum_{i=1}^{k+1} C(n - k, i)e^{-(\mu + i\lambda a)t}, 1 \leq k \leq n - 1 \tag{12}$$

where $C(n - k, i)$, $1 \leq k \leq n - 1, 0 \leq i \leq k + 1$, are real constants determined by the initial condition of the ODEs.

Finally, substituting (12) into (5) and solving the ODE, we have

$$p_0(t) = C(0, 0) + \sum_{i=1}^n C(0, i)e^{-(\mu + i\lambda a)t} + C(0, n + 1)e^{-(n+1)\lambda a t} \tag{13}$$

where $C(0, i)$, $0 \leq i \leq n + 1$, are real constants determined by the initial condition of the ODEs.

Equations (9), (12) and (13) consist of the transient solution of the queueing system. The steady state system solution can be obtained when $t \rightarrow \infty$. Thus, the steady state probabilities are $p_k = C(k, 0), 0 \leq k \leq n$.

In practice, the queueing system is usually empty in the beginning. Thus, the initial condition of the ODEs is:

$$p_0(0) = 1, p_i(0) = 0, 1 \leq i \leq n. \tag{14}$$

By applying the initial condition of the system (14) to the transient solution of Equations (9), (12) and (13), we can recursively determine all the coefficients $C(n - k, i), 0 \leq k \leq n, 0 \leq i \leq k + 1$.

$$C(n, 0) = \frac{\lambda a}{\mu + \lambda a} \tag{15}$$

$$C(n, 1) = -\frac{\lambda a}{\mu + \lambda a} \tag{16}$$

$$C(n - k, 0) = \frac{1}{\mu + (k + 1)\lambda a} \left[\lambda a + \mu C(n - k + 1, 0) - \lambda a \sum_{i=1}^k C(n - k + i, 0) \right], 1 \leq k < n \tag{17}$$

$$C(n - k, j) = \frac{1}{(k + 1 - j)\lambda a} \left[\mu C(n - k + 1, j) - \lambda a \sum_{i=1}^{k+1-j} C(n - k + i, j) \right], \quad 1 \leq k < n, 1 \leq j \leq k \tag{18}$$

$$C(n - k, k + 1) = - \sum_{j=0}^k C(n - k, j), \tag{19}$$

$$C(0, 0) = \frac{\mu C(1, 0)}{n\lambda a} \tag{20}$$

$$C(0, j) = \frac{\mu C(1, j)}{-\mu + (n - j)\lambda a}, \quad 1 \leq j \leq n \tag{21}$$

$$C(0, n + 1) = 1 - \sum_{j=0}^n C(0, j), \tag{22}$$

It can be easily shown that $C(0, n + 1)$ is equal to zero by substituting related expressions into Equation (22). After all the coefficients are solved, the transient solution to the queueing model will be obtained. The steady state solution to the model will also be obtained by letting t approach infinity.

4. Case Study

In this section, we present a case study for the proposed queueing model with $n = 4$ (i.e., one server and three waiting places). The queueing system can be re-written as follows.

$$p'_0(t) = -4\lambda a p_0(t) + \mu p_1(t), \tag{23}$$

$$p'_1(t) = -[\mu + 3\lambda a]p_1(t) + \lambda a \left[1 - \sum_{i=1}^4 p_i(t) \right] + \mu p_2(t), \tag{24}$$

$$p'_2(t) = -[\mu + 2\lambda a]p_2(t) + \lambda a \left[1 - \sum_{i=2}^4 p_i(t) \right] + \mu p_3(t), \tag{25}$$

$$p'_3(t) = -[\mu + \lambda a]p_3(t) + \lambda a \left[1 - \sum_{i=3}^4 p_i(t) \right] + \mu p_4(t), \tag{26}$$

$$p'_4(t) = -\mu p_4(t) + \lambda a [1 - p_4(t)]. \tag{27}$$

The initial condition is:

$$p_0(0) = 1, \quad p_i(0) = 0, \quad 1 \leq i \leq 4. \tag{28}$$

The above queueing system can be solved as the Initial Value Problem (IVP) of ODEs. The solution is as follows.

$$p_4(t) = C(4, 0) + C(4, 1)e^{-(\mu + \lambda a)t} \tag{29}$$

$$p_3(t) = C(3, 0) + C(3, 1)e^{-(\mu + \lambda a)t} + C(3, 2)e^{-(\mu + 2\lambda a)t} \tag{30}$$

$$p_2(t) = C(2, 0) + C(2, 1)e^{-(\mu + \lambda a)t} + C(2, 2)e^{-(\mu + 2\lambda a)t} + C(2, 3)e^{-(\mu + 3\lambda a)t}, \tag{31}$$

$$p_1(t) = C(1, 0) + C(1, 1)e^{-(\mu + \lambda a)t} + C(1, 2)e^{-(\mu + 2\lambda a)t} + C(1, 3)e^{-(\mu + 3\lambda a)t} + C(1, 4)e^{-(\mu + 4\lambda a)t}, \tag{32}$$

$$p_0(t) = C(0, 0) + C(0, 1)e^{-(\mu + \lambda a)t} + C(0, 2)e^{-(\mu + 2\lambda a)t} + C(0, 3)e^{-(\mu + 3\lambda a)t} + C(0, 4)e^{-(\mu + 4\lambda a)t}, \tag{33}$$

where the coefficients are

$$C(4, 0) = \frac{\lambda a}{\mu + \lambda a} \tag{34}$$

$$C(4, 1) = -\frac{\lambda a}{\mu + \lambda a} \tag{35}$$

$$C(3,0) = \frac{2\mu\lambda a}{(\mu + \lambda a)(\mu + 2\lambda a)} \tag{36}$$

$$C(3,1) = -\frac{\mu - \lambda a}{\mu + \lambda a} \tag{37}$$

$$C(3,2) = \frac{\mu - 2\lambda a}{\mu + 2\lambda a} \tag{38}$$

$$C(2,0) = \frac{3\mu^2\lambda a}{(\mu + \lambda a)(\mu + 2\lambda a)(\mu + 3\lambda a)} \tag{39}$$

$$C(2,1) = -\frac{\mu(\mu - 2\lambda a)}{2\lambda a(\mu + \lambda a)} \tag{40}$$

$$C(2,2) = \frac{(\mu - \lambda a)(\mu - 2\lambda a)}{\lambda a(\mu + 2\lambda a)} \tag{41}$$

$$C(3,1) = -\frac{\mu - \lambda a}{\mu + \lambda a} \tag{42}$$

$$C(3,2) = \frac{\mu - 2\lambda a}{\mu + 2\lambda a} \tag{43}$$

$$C(2,0) = \frac{3\mu^2\lambda a}{(\mu + \lambda a)(\mu + 2\lambda a)(\mu + 3\lambda a)} \tag{44}$$

$$C(2,1) = -\frac{\mu(\mu - 2\lambda a)}{2\lambda a(\mu + \lambda a)} \tag{45}$$

$$C(2,2) = \frac{(\mu - \lambda a)(\mu - 2\lambda a)}{\lambda a(\mu + 2\lambda a)} \tag{46}$$

$$C(2,3) = \frac{-\mu^3 + 2\mu^2\lambda a + 2\mu\lambda^2 a^2 - 12\lambda^3 a^3}{2\lambda a(\mu + 2\lambda a)(\mu + 3\lambda a)} \tag{47}$$

$$C(1,0) = \frac{4\mu^3\lambda a}{(\mu + \lambda a)(\mu + 2\lambda a)(\mu + 3\lambda a)(\mu + 4\lambda a)} \tag{48}$$

$$C(1,1) = -\frac{\mu^2(\mu - 3\lambda a)}{6\lambda^2 a^2(\mu + \lambda a)} \tag{49}$$

$$C(1,2) = -\frac{\mu(\mu - 2\lambda a)^2}{2\lambda^2 a^2(\mu + 2\lambda a)} \tag{50}$$

$$C(1,3) = \frac{-\mu^4 + 3\mu^3\lambda a - 14\mu\lambda^3 a^3 + 12\lambda^4 a^4}{2\lambda^2 a^2(\mu + 2\lambda a)(\mu + 3\lambda a)} \tag{51}$$

$$C(1,4) = \frac{\mu^5 - \mu^4\lambda a - 6\mu^3\lambda^2 a^2 + 30\mu^2\lambda^3 a^3 - 12\mu\lambda^4 a^4 - 144\lambda^5 a^5}{6\lambda^2 a^2(\mu + 2\lambda a)(\mu + 3\lambda a)(\mu + 4\lambda a)} \tag{52}$$

$$C(0,0) = \frac{\mu^4}{(\mu + \lambda a)(\mu + 2\lambda a)(\mu + 3\lambda a)(\mu + 4\lambda a)} \tag{53}$$

$$C(0,1) = -\frac{\mu^3}{6\lambda^2 a^2(\mu + \lambda a)} \tag{54}$$

$$C(0,2) = -\frac{\mu^2(\mu - 2\lambda a)}{2\lambda^2 a^2(\mu + 2\lambda a)} \tag{55}$$

$$C(0,3) = \frac{\mu(\mu^3 - 2\mu^2\lambda a - 2\mu\lambda^2 a^2 + 12\lambda^3 a^3)}{2\lambda^2 a^2(\mu + 2\lambda a)(\mu + 3\lambda a)} \tag{56}$$

$$C(0,4) = -C(1,4). \tag{57}$$

It can be verified that the solution to the case study is consistent with the solution to the general queueing model described by Equations (9), (12) and (13) along with Equations (15) to (22). The corresponding steady state solution is given by $p_k = C(k, 0)$, $0 \leq k \leq 4$.

5. Performance Metrics

5.1. Blocking Probability of the System

The system blocking probability, denoted by $P_B(t)$, is equal to the probability that there are n customers in the system including the one in service, so the new arrival of customer has to be blocked. Thus, we have

$$P_B(t) = p_n(t) = C(n, 0) + C(n, 1)e^{-(\mu+\lambda a)t} \quad (58)$$

5.2. Availability of the System

The availability of the system, denoted by $A_S(t)$, is defined as the probability that the system can accept at least one customer.

$$A_S(t) = 1 - p_n(t) \quad (59)$$

5.3. Idle and Busy Probabilities of the System

The system idle probability, denoted by $P_{Id}(t)$, is defined as the probability that the system is empty (i.e., no customers in service and in the queue).

$$P_{Id}(t) = p_0(t). \quad (60)$$

The system busy probability, denoted by $P_{Bu}(t)$, is defined as the probability that there is at least one customer in the system.

$$P_{Bu}(t) = 1 - p_0(t). \quad (61)$$

5.4. Queueing Probability of the System

The system queueing probability, denoted by $P_Q(t)$, is defined as the probability that there is at least one customer in the queue.

$$P_Q(t) = \sum_{k=2}^n p_k(t) \quad (62)$$

5.5. Mean Number of Customers in the System and Queue

The mean number of customers in the system, $N_S(t)$, can be expressed as

$$L_S(t) = \sum_{k=1}^n k p_k(t) \quad (63)$$

Similarly, the mean number of customers in the queue, $N_Q(t)$, can be calculated as

$$L_Q(t) = \sum_{k=2}^n (k-1) p_k(t) \quad (64)$$

5.6. Mean Waiting Time of Customers in the System and Queue

By Little's law, the mean waiting time of customers in the system, denoted by $W_S(t)$, can be determined as

$$W_S(t) = \frac{L_S(t)}{\lambda'} \quad (65)$$

where λ' is the effective arrival rate, i.e., the mean rate of the customers actually entering the system. As the system is finite and the arriving customers who find the system full leave the system, the effective arrival rate can be calculated as

$$\lambda' = \lambda[1 - P_B(t)] \quad (66)$$

Similarly, the mean waiting time of customers in the queue, denoted by $W_Q(t)$, can be determined as

$$W_Q(t) = \frac{L_Q(t)}{\lambda'} \tag{67}$$

6. Numerical Results

After obtaining the transient analysis (and thus the steady state analysis) of the finite queueing model and the developed performance metrics, we can apply them to many different fields such as healthcare, telecommunication, commerce, manufacturing, transportation, etc. However, we will leave these to our future study. Rather, in this section we present numerical results for the validation of our analytical solution under different parameters. The typical parameter settings are given in Table 1.

Table 1. Typical parameter configuration for numerical evaluation.

Parameters	Value	Unit	Description
t	[0, 30]	General time units	time
λ	0.4, 0.9	Customers/unit time	arrival rate
μ	0.5, 1.0	Customers/unit time	service rate
n	10	Number of customers	system capacity
$a_i = a, i \leq n$	0.1		Prob. distribution of bulk arrivals

Figure 2 shows how the transient solution of the blocking probability $P_B(t)$ changes with respect to time t and the arrival and service rates λ and μ . The probability $P_B(t)$ will gradually increase and then tend to stabilize as the time goes. This is reasonable as more places are initially available for the arriving customers and gradually become less and less. We also observe that $P_B(t)$ will increase when the arrival rate (λ) or the service time ($1/\mu$) is increased. The increase of the arrival rate will lead to more customers entering the queue and thus cause the system to tend to block. Equivalently, the increase of the service time will cause more customers to wait in the queue.

Figure 3 shows the transient solution of the system availability $A_S(t)$ with respect to time t and the arrival and service rates λ and μ . The availability $A_S(t)$ will gradually decrease and then tend to be constant with respect to time. The system will have more customers and thus less available as time goes. We also observe that $A_S(t)$ will decrease when the arrival rate (λ) or the service time ($1/\mu$) is increased. The more customers in the system, the less availability for the system.

Figure 4 shows the busy probability of the system $P_{Bu}(t)$ with respect to time and other parameters. As expected, $P_{Bu}(t)$ will increase with respect to time as more and more customers arrive at the system and the system becomes busier. Similarly, when the arrival rate increases or the service rate reduces, $P_{Bu}(t)$ will become busier.

Figure 5 shows the queueing probability of the system $P_Q(t)$ with respect to time and other parameters. Similar to $P_{Bu}(t)$, we observe that $P_Q(t)$ is zero in the beginning of time (i.e., $t = 0$) and will increase with respect to time. We also observe that $P_Q(t)$ will increase when λ is increased or μ is decreased. More customer arrivals or less service capability of the service station leads to a large queueing probability.

Figure 6 shows the mean number of customers in the system $L_S(t)$ with respect to time and arrival and service rates. We observe that $L_S(t)$ will increase with the increase of the arrival rate λ and the service time ($1/\mu$). In a single server queueing system, it is obvious that the increase of the customer arrival rate will lead to the increase of the queueing length. Equivalently, the increase of the customer service time will cause more customers to wait in the queue and thus lead to the increase of queueing length. Similarly, the mean number of customers in the queue $L_Q(t)$ has the same characteristics, as shown in Figure 7.

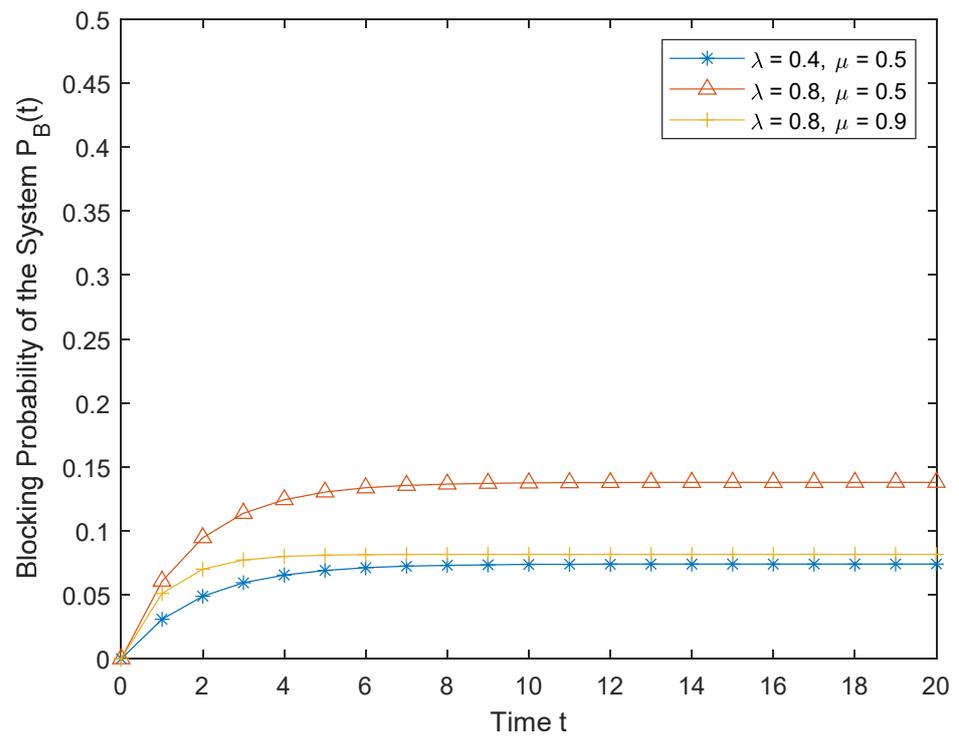


Figure 2. The blocking probability of the system $P_B(t)$.

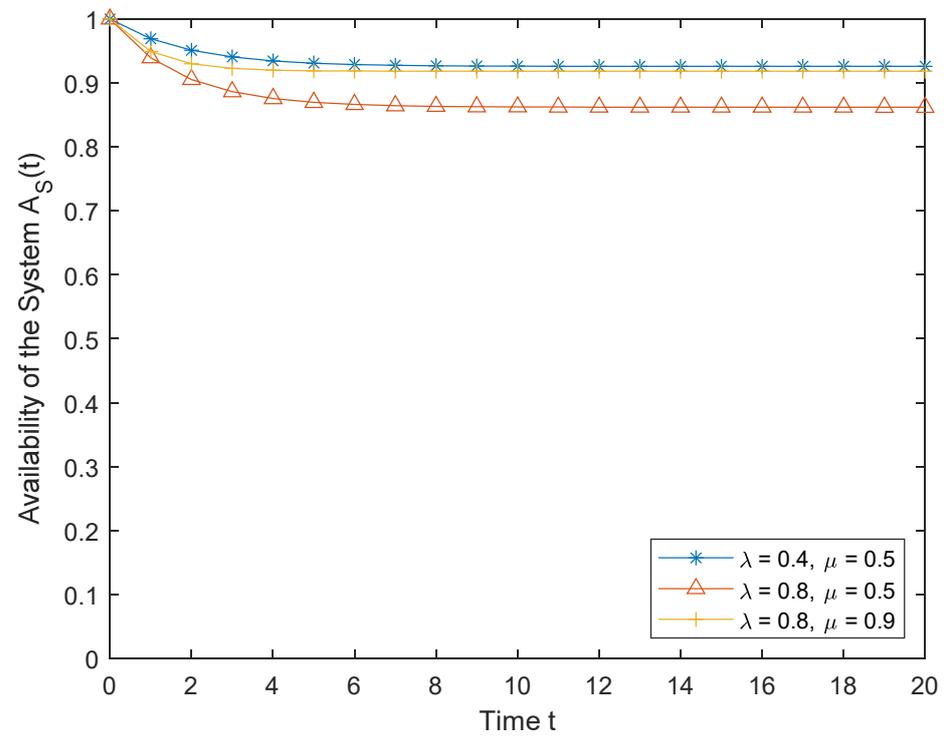


Figure 3. The system availability $A_S(t)$.

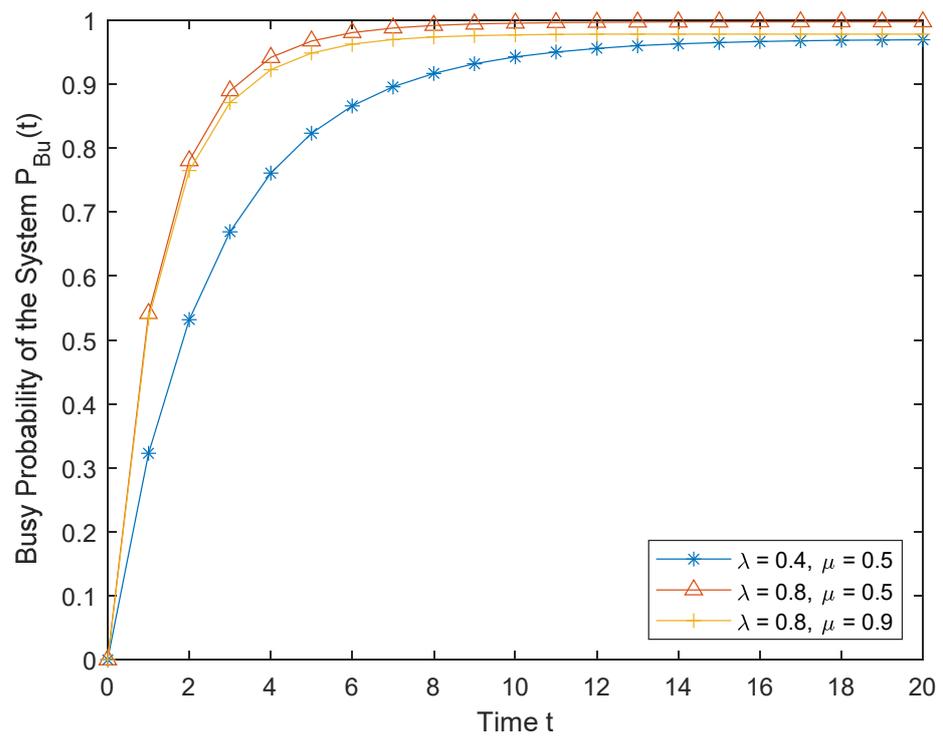


Figure 4. The busy probability of the system $P_{Bu}(t)$.

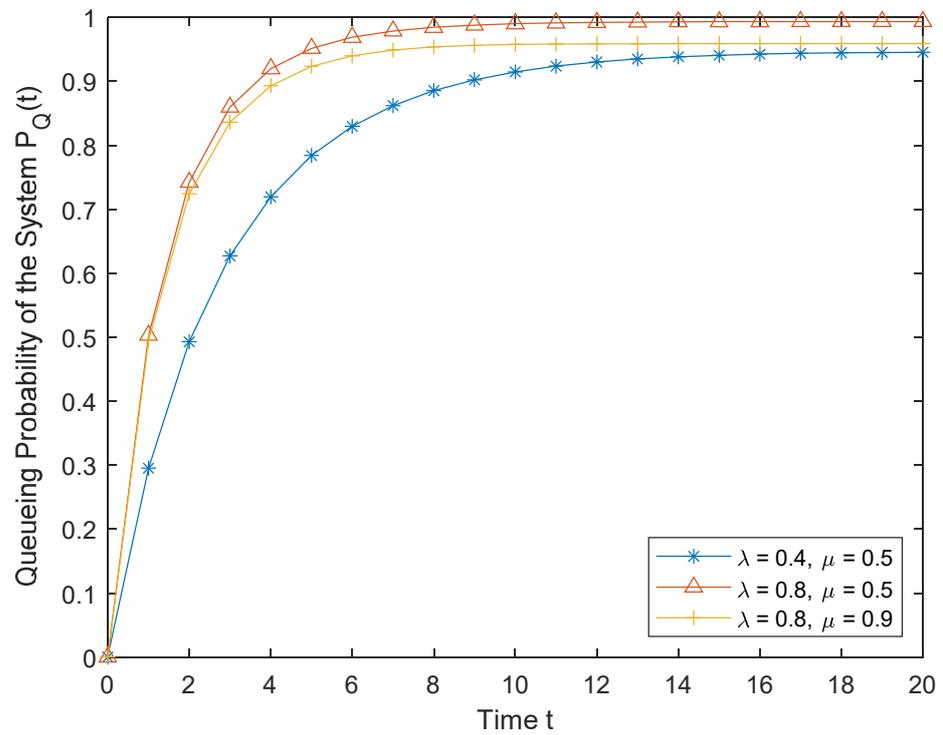


Figure 5. The queuing probability of the system $P_Q(t)$.

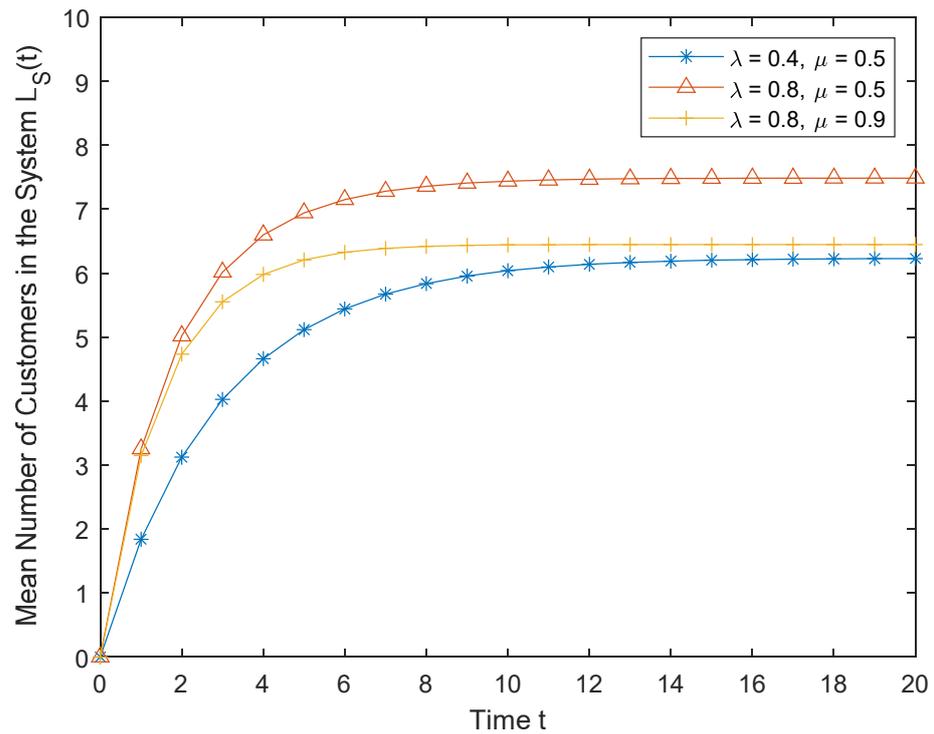


Figure 6. The mean number of customers in the system $L_S(t)$.

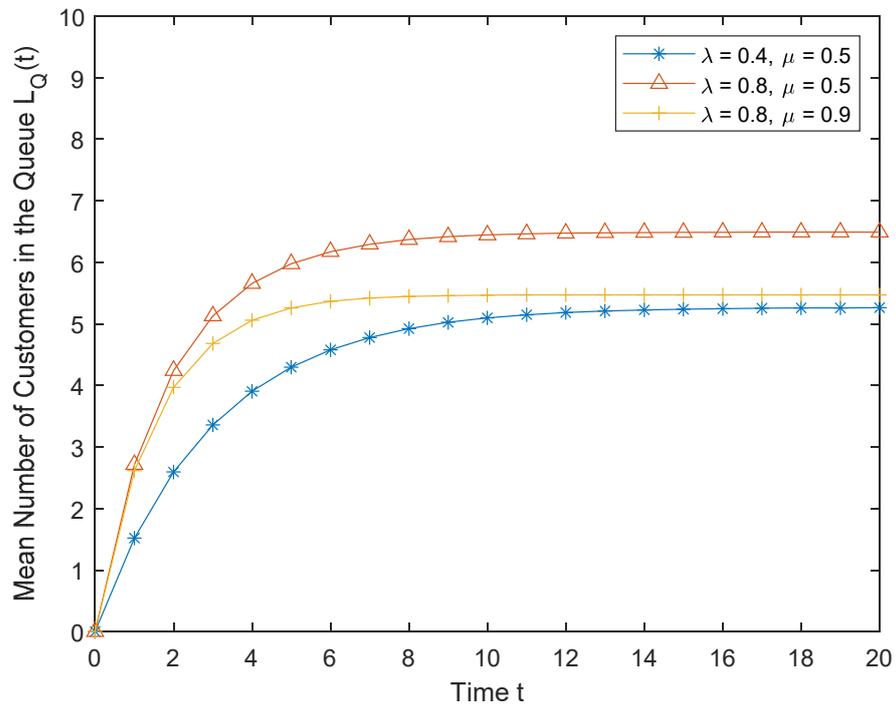


Figure 7. The mean number of customers in the queue $L_Q(t)$.

Figure 8 shows the mean waiting time of customers in the system $W_S(t)$ with respect to the change of time and service rate. We observe that $W_S(t)$ will increase with respect to time initially and then tends to stabilize as the time goes. We also observe that $W_S(t)$ will increase when the service rate is decreased. The decrease of service rate means the increase of service time per customer, leading to the increase of waiting time of other customers. The mean waiting time of customers in the queue $W_Q(t)$ has the same trend as $W_S(t)$ except with shorter time, as shown in Figure 9.

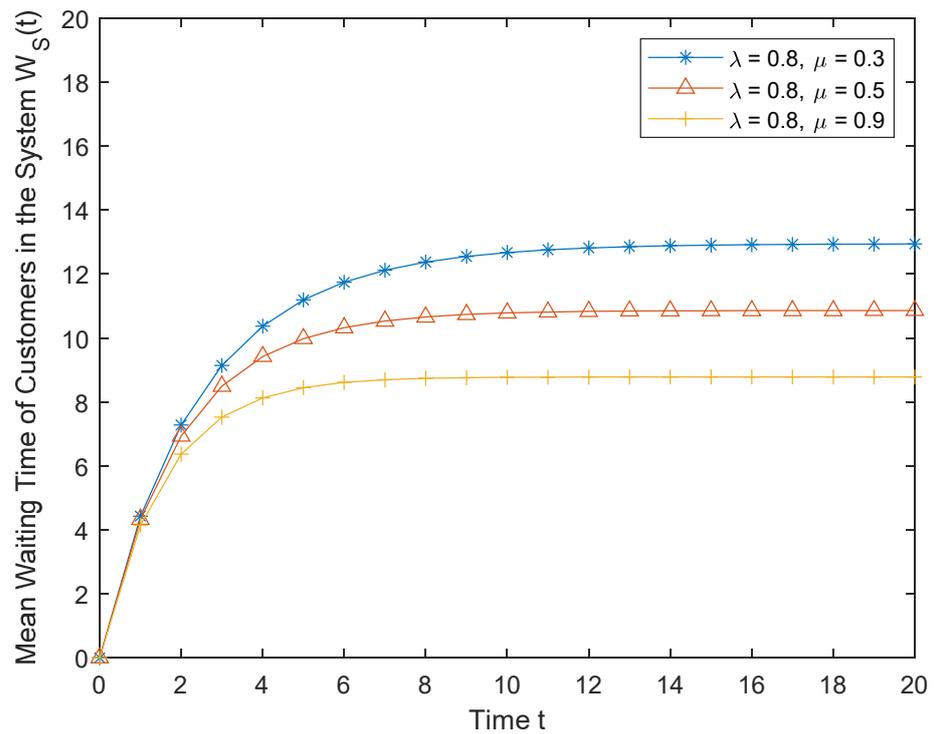


Figure 8. The mean waiting time of customers in the system $W_S(t)$.

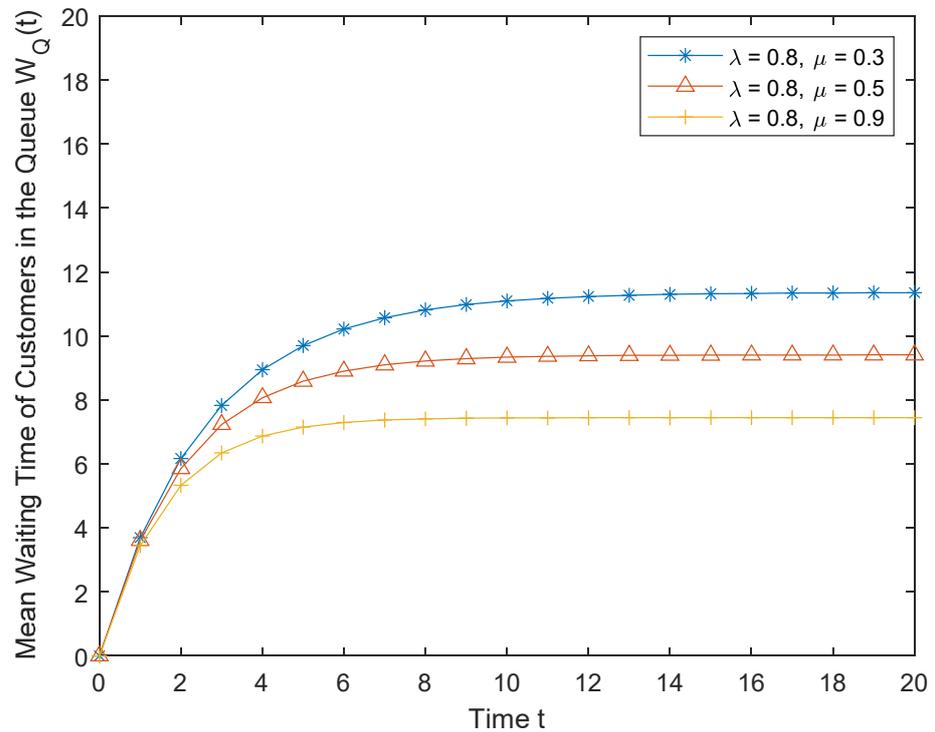


Figure 9. The mean waiting time of customers in the queue $W_Q(t)$.

7. Conclusions

We proposed a finite capacity single server queueing model with bulk arrivals and performed the transient analysis for the system based on different system parameters. By using the ordinary differential equations (ODEs) technique, we derived the transient solution in explicit form for the queueing system and thus the steady state solution. A case study of the queueing model was presented for validation of the obtained transient

solution. Some performance metrics of interest were then developed along with numerical results for further understanding. Our numerical results validated the analytical solutions of all the performance metrics including the transient solution of the blocking probability, the transient solution of the system availability, the busy probability of the system, the queueing probability of the system, the mean number of customers in the system and in the queue, and the mean waiting time of customers in the system and in the queue (Please refer the details in Section 6). Although the queueing model is initiated from the IoT based edge computing platform, the proposed system modeling and analysis method can be applied to a wider range of applications such as healthcare, telecommunication, commerce, manufacturing, transportation, etc.

Funding: This work was supported in part by the Faculty Improvement Grant (No. 211208) from Inter Faculty Organization (IFO)/Minnesota State Master Agreement, MN, USA.

Data Availability Statement: Not applicable.

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

Nomenclature

n	The total number of customers in the system.
$k, 0 \leq k \leq n$	The current number of customers in the system (i.e., the system state).
λ	The mean arrival rate of the bulk arrivals.
μ	The mean service rate of the single server system.
X	The random variable used to describe the number of arrivals in a group (i.e., batch size).
$a_i, 1 \leq i \leq n$	The probability distribution of batch size, $a_i = P(X = i)$.
a	The probability of the batch size when the batch size X is a uniform random variable.
$p_k(t), 0 \leq k \leq n$	The probability that k customers are present in the system at time t .
$C(k, i)$	The coefficients of the solution to the ODEs.
λ'	The effective arrival rate.
$P_B(t)$	The blocking probability of the system.
$A_S(t)$	The system availability.
$P_{id}(t)$	The system idle probability.
$P_{Bu}(t)$	The system busy probability.
$P_Q(t)$	The system queueing probability.
$L_S(t)$	The mean number of customers in the system.
$L_Q(t)$	The mean number of customers in the queue.
$W_S(t)$	The mean waiting time of customers in the system.
$W_Q(t)$	The mean waiting time of customers in the queue.

References

1. Kleinrock, L. *Queueing Systems, Volume 1: Theory*, 1st ed.; Wiley-Interscience: Hoboken, NJ, USA, 1975.
2. Chaudhry, M.L.; Templeton, J.G.C. *A First Course in Bulk Queues*; John Wiley & Sons: Hoboken, NJ, USA, 1983; ISBN 978-0471862604.
3. Pegden, C.D.; Rosenshine, M. Some New Results for the M/M/1 Queue. *Manag. Sci.* **1982**, *28*, 821–828. [[CrossRef](#)]
4. Lopez-Herrero, M.J. Distribution of the Number of Customers Served in an M/G/1 Retrial Queue. *J. Appl. Probab.* **2002**, *39*, 407–412. [[CrossRef](#)]
5. Conolly, B.W. Queueing at a Single Serving Point with Group Arrival. *J. R. Stat. Soc. Ser. B* **1960**, *22*, 285–298. [[CrossRef](#)]
6. Shanbhag, D.N. On Infinite Server Queues with Batch Arrivals. *J. Appl. Probab.* **1966**, *3*, 274–279. [[CrossRef](#)]
7. Sharma, S.D. On a Continuous/Discrete Time Queueing System with Arrivals in Batches of Variable Size and Correlated Departures. *J. Appl. Probab.* **1975**, *12*, 115–129. [[CrossRef](#)]
8. Alfa, A.S. A Numerical Method for Evaluating Delay to a Customer in a Time-Inhomogeneous, Single Server Queue with Batch Arrivals. *J. Oper. Res. Soc.* **1979**, *30*, 665–667. [[CrossRef](#)]
9. van Dijk, N.M. An LCFS Finite Buffer Model with Finite Source Batch Input. *J. Appl. Probab.* **1989**, *26*, 372–380. [[CrossRef](#)]
10. Choi, B.D.; Park, K.K. The Mk/m/∞ Queue with Heterogeneous Customers in a Batch. *J. Appl. Probab.* **1992**, *29*, 477–481.
11. Chen, S.-P. A bulk arrival queueing model with fuzzy parameters and varying batch sizes. *Appl. Math. Model.* **2006**, *30*, 920–929. [[CrossRef](#)]
12. Baruah, M.; Madan, K.C.; Eldabi, T. A Batch Arrival Single Server Queue with Server Providing General Service in Two Fluctuating Modes and Reneging during Vacation and Breakdowns. *J. Probab. Stat.* **2014**, *2014*, 319318. [[CrossRef](#)]
13. Niranjana, S.P.; Indhira, K.; Chandrasekaran, V.M. Analysis of bulk arrival queueing system with batch size dependent service and working vacation. *AIP Conf. Proc.* **2018**, *1952*, 020061. [[CrossRef](#)]
14. Laslett, G.M. Characterising the Finite Capacity GI/M/1 Queue with Renewal Output. *Manag. Sci.* **1975**, *22*, 106–110. [[CrossRef](#)]

15. Chao, X. On the Departure Processes of M/M/1/N and GI/G/1/N Queues. *Adv. Appl. Probab.* **1992**, *24*, 751–754. [[CrossRef](#)]
16. Kaczynski, W.H.; Leemis, L.M.; Drew, J.H. Transient Queueing Analysis. *INFORMS J. Comput.* **2012**, *24*, 10–28. [[CrossRef](#)]
17. Kelton, W.D.; Law, A.M. The Transient Behavior of the M/M/s Queue, with Implications for Steady-State Simulation. *Oper. Res.* **1985**, *33*, 378–396. [[CrossRef](#)]
18. Sharma, O.P.; Gupta, U.C. Transient Behaviour of an M/M/1/N queue. *Stoch. Process. Appl.* **1982**, *13*, 327–331. [[CrossRef](#)]
19. Leguesdron, P.; Pellaumail, J.; Rubino, G.; Sericola, B. Transient Analysis of the M/M/1 Queue. *Adv. Appl. Probab.* **1993**, *25*, 702–713. [[CrossRef](#)]
20. Sharmaand, O.P.; Tarabia, A.M.K. A Simple Transient Analysis of an M/M/1/N Queue. *Sankhyā Indian J. Stat. Ser. A* **2000**, *62*, 273–281.
21. Garcia, J.-M.; Brun, O.; Gauchard, D. Transient Analytical Solution of M/D/1/N Queues. *J. Appl. Probab.* **2002**, *39*, 853–864. [[CrossRef](#)]
22. Kaczynski, W.; Leemis, L.; Drew, J. Modeling and analyzing transient military air traffic control. In Proceedings of the 2010 Winter Simulation Conference, Baltimore, MD, USA, 5–8 December 2010; pp. 1395–1406. [[CrossRef](#)]
23. Chydzinski, A.; Adamczyk, B. Transient and Stationary Losses in a Finite-Buffer Queue with Batch Arrivals. *Math. Probl. Eng.* **2012**, *2012*, 326830. [[CrossRef](#)]
24. Kreyszig, E. *Advanced Engineering Mathematics*, 10th ed.; Wiley: Hoboken, NJ, USA, 2020; ISBN 978-1119455929.