

Article

Optimal Service Time Distribution for an $M/G/1$ Waiting Queue

Mario Lefebvre *  and Roozbeh Yaghoubi 

Department of Mathematics and Industrial Engineering, Polytechnique Montréal, C.P. 6079, Succursale Centre-ville, Montréal, QC H3C 3A7, Canada; roozbeh.yaghoubi@polymtl.ca

* Correspondence: mlefebvre@polymtl.ca

Abstract: Suppose that in an $M/G/1$ waiting queue, the server can choose between two service time distributions. We look for the choice that enables us to minimize the expected value of a cost criterion that takes into account the cost incurred by working faster and the time needed to empty the waiting line. The random final time is the first time there is no customer waiting for service. When the service times are exponential random variables, we can appeal to dynamic programming to obtain the optimal solution. In the general case, conditional probability is used. Particular problems in which the capacity of the system is finite are solved explicitly.

Keywords: $M/M/1$ model; active queue management; service rate policy; dynamic programming; first-passage time; homing problem

MSC: Primary 93E20; Secondary 60K30



Citation: Lefebvre, M.; Yaghoubi, R. Optimal Service Time Distribution for an $M/G/1$ Waiting Queue. *Axioms* **2024**, *13*, 594. <https://doi.org/10.3390/axioms13090594>

Academic Editors: Santiago E. Moll, Luis Manuel Sánchez Ruiz and Behzad Djafari-Rouhani

Received: 13 June 2024

Revised: 21 August 2024

Accepted: 30 August 2024

Published: 31 August 2024



Copyright: © 2024 by the authors. 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

Queuing theory plays an important role in decision making by providing insight into the operation and efficiency of queuing systems. It deals with the mathematical modeling of customer arrivals and the corresponding service times in queuing systems. A key aspect of queuing theory is decision optimization, which focuses on examining the underlying parameters and constraints of a model and helps decision makers optimize performance measures. Various queue configurations can be analyzed, such as the speed of service and number of servers, to assess the most cost-effective and efficient strategy.

Another aspect of queuing theory in the context of decision making is sensitivity analysis. It allows decision-makers to analyze various scenarios by modeling several queue configurations and to identify the impact of variations in system parameters, particularly the cost of changes in customer service models. In addition, it allows decision makers to realize the possible consequences of different system parameters and to take informed action by understanding system dynamics (see [1]).

Queues are used in a variety contexts, such as mobile data in the wireless mobile systems discussed by Li et al. [2] and patient flow in the clinics studied by Aziati and Hamdan [3], etc.

Reducing customer congestion in any service system is a common problem, and it is the main challenge in queue management. Its importance is highlighted in a variety of ways in numerous research studies, such as, for example, the loss of productivity that occurs due to time lost in traffic (De and Rajbongshi [4]), health impacts (Requia et al. [5]), etc.

To solve the problem of reducing customer congestion in queues, increasing the number of servers assigned to a system is an immediate solution, but adding each server incurs service costs. This challenge has been taken up by many researchers in order to determine the optimal number of servers while considering the cost function as an objective, as explained by Lefebvre and Yaghoubi [6]. Consequently, a compromise must be found between customer congestion and the total cost of a faster service, either through the speed

of service or the number of servers. Therefore, characterizing the optimal-service rate in a system is a fundamental factor in practice for how operations should be conducted.

Many papers have investigated variations in the service rate to reduce queues. In the age of artificial intelligence, a wide range of optimization techniques based on computational methods have been introduced, as shown by Laxmi and Jyothsna [7]. The aforementioned authors examined a discrete-time queue using particle swarm optimization to optimize service rates in order to minimize costs, and their paper explored its application in inbound email contact centers.

New mathematical programming models are recent directions of the research in this area, such as Tian et al. [8]. These authors studied a single-server queuing system with varying service rates to analyze how customer decisions affect the equilibrium performance of a queue using continuous-time Markov chains for modeling and analysis. Singh et al. [9] investigated a method for detecting change points in the service rate of an $M/M/1/m$ queue by analyzing the changes in the probability distribution of a stochastic process using a likelihood function based on the number of customers observed at departures and via employing Bayesian estimators.

Laxmi et al. [10] proposed an $M/M/2$ queuing system with an alternative second service using a system stabilization and performance evaluation method. Wu et al. [11] examined a queuing system that could be used to control customer arrivals, as well as a $GI/M/1$ queue with variable service speeds. They performed a steady-state analysis for system optimization, with the objective of reducing costs and waiting times. The results show a diminishing return on additional investment in efficient systems, and regression models are proposed for decision making. Chen et al. [12] studied service rate control using event-based optimization to obtain optimal control policies. They proposed an algorithm for determining service rates and compared its effectiveness with state-based optimization.

Dudin et al. [13] analyzed a queuing system with two different types of requests: Type 1, i.e., with a constant service rate; and Type 2, i.e., with flexible rates. They explored the behavior of the system through a Markov chain, focusing on performance measures and the optimal service assigned for the two types of requests. Lakkumikanthan and Balasubramanian [14] developed an optimal service rate decision for a discrete-time queuing system with a finite buffer employing Markov decision theory and linear programming. The aim was to optimize costs by regulating service rates based on customer arrival and inventory reloading. Su and Li [15] focused on a single-server queue in communication networks with varying service rates and various types of customers. Their paper compared control levels based on various parameters and chose an optimal policy when multiple optimal policies were available. Büke and Qin [16] studied multi-server queuing systems with varying rates of service using stochastic processes to analyze performance. They used martingales to characterize these processes under different policies.

For an overview of other studies in this area, the reader is referred to Dai et al. [17], Chen and Xia [18], and Rodrigues et al. [19].

In this paper, we suppose that the random service time is either S_1 or S_2 . Moreover, in Section 2, the server can switch between S_1 and S_2 at any time instant.

2. Problem Description

We denote respectively by $X(t)$ and $X_q(t)$ the total number of customers in the system and the number of customers who are waiting in line at time t . We suppose that at the initial time t_0 , $X(t_0) = X_q(t_0) \in \{2, 3, \dots, c_p\}$, where c_p is the capacity of the system, which can be infinite. The server therefore starts his/her service at time t_0 , which could be, for example, the opening time of a shop.

We define the *first-passage time* $T(k) [= T(k, t_0, t_1)]$ as follows:

$$T(k) = \inf\{t > t_0 : X_q(t) = 0 \text{ or } t = t_1 (> t_0) \mid X(t_0) = k\}. \quad (1)$$

Our aim is to determine whether the server should choose the service time S_1 or S_2 at time t to minimize the expected value of the criterion function:

$$J(k) := \int_{t_0}^{T(k)} \left\{ \frac{q_0}{(E[S(t)])^2} + \lambda \right\} dt + K[T(k)], \quad (2)$$

where $S(t) = S_1$ or S_2 , q_0 and λ are positive constants, and $K(\cdot)$ is the terminal cost function. If $E[S_1] > E[S_2]$, then the instantaneous (or running) cost $q_0/(E[S(t)])^2$ will be smaller. However, it will take longer to reduce the number of waiting customers to zero, thus incurring a larger penalty (since $\lambda > 0$). The optimizer must also take into account the terminal cost $K[T(k)]$.

Remark 1. The term $(E[S(t)])^2$ in the cost function $J(k)$ corresponds to $u^2(t)$ in LQ or LQG optimal control problems. Moreover, when $S(t)$ does not depend on t and $E[T(k)] = E[S]$, as will be the case in Section 4, if we use $q_0/E[S]$ in the cost function, then $E[S]$ cancels out, so that the expected service time is not taken into account.

The optimal control problem considered is a particular *homing problem* (see Whittle [20]). In these problems, the optimizer tries to find the control of a stochastic process that minimizes the expected value of a given cost function between an initial time and the time when a certain event occurs. Whittle [20] treated the case when the stochastic process is an n -dimensional diffusion process. Rishel [21] set up such problems for n -dimensional diffusion processes that can be used to model the wear of devices. Lefebvre has written numerous papers on homing problems. In a recent one ([22]), he extended homing problems to the case of autoregressive processes (see also Lefebvre and Pazhoheshfar [23] for an optimal maintenance problem).

Homing problems for modified $M/G/k$ queuing systems were treated by Lefebvre in [24,25], and by Lefebvre and Yaghoubi [6]. In these papers, the control variable was the number of servers working at time t .

Remark 2. There are numerous applications of queuing theory. In many situations, time $T(k)$ is indeed a random variable. For example, suppose that the “customers” in the systems are orders received by a company. The time required to fill the orders is generally not fixed. Moreover, the constant t_1 could be the deadline by which orders are fulfilled, where otherwise a penalty is imposed on the company.

In Section 3, we will present an aeronautical application. In this case, the customers are aircraft that have been brought to a workshop for maintenance. The aim, of course, is to get the aircraft back into service as quickly as possible. Once again, the time required to carry out maintenance is random since problems may be detected during the basic inspection.

Next, we define the *value function*

$$F(k, t_0) = \inf_{\substack{S(t) \\ t \in [t_0, T(k)]}} E[J(k)]. \quad (3)$$

That is, $F(k, t_0)$ is the expected cost incurred if the optimizer chooses the optimal control between the initial time t_0 and the random final time $T(k)$. Moreover, we assume that

$$K[T(k)] = \begin{cases} 0 & \text{if } T(k) \in [t_0, t_1), \\ F_1 & \text{if } T(k) = t_1, \end{cases} \quad (4)$$

where $F_1 > 0$. It then follows that the value function satisfies the boundary conditions

$$F(k, t_0) = \begin{cases} 0 & \text{if } k = 1, \\ F_1 & \text{if } (k > 1 \text{ and}) t_0 = t_1. \end{cases} \quad (5)$$

Remark 3. The aim is to empty the queue at the latest by time t_1 , which is a fixed constant. If we have not succeeded in doing so by time t_1 , we stop the optimization problem and a cost F_1 is incurred.

In Section 3, we will use the fact that, when $S_i \sim \text{Exp}(\mu_i)$, that is, when S_i has an exponential distribution with parameter μ_i for $i = 1, 2$, it is possible to make use of dynamic programming to determine the optimal control. In the general case, assuming that the optimizer chooses either S_1 or S_2 for all t in $[t_0, T(k))$, we will use conditional probability to find the optimal solution. This will be conducted in Section 4. Finally, we will conclude with a few remarks in Section 5.

3. Optimal Control of the Modified M/M/1 Queue

When $S_i \sim \text{Exp}(\mu_i)$, we have a modified M/M/1 queueing model, in which the service time distribution is not fixed. We can then write the cost function $J(k)$ as follows:

$$J(k) = \int_{t_0}^{T(k)} \{q_0 \mu^2(t) + \lambda\} dt + K[T(k)], \quad (6)$$

where $\mu(t) = \mu_1$ or μ_2 .

Let A be the time needed for a new customer to arrive after time t_0 . We have $A \sim \text{Exp}(\lambda)$, so that

$$P[A \leq \Delta t] = 1 - e^{-\lambda \Delta t} = \lambda \Delta t + o(\Delta t), \quad (7)$$

and, similarly,

$$P[S_i \leq \Delta t] = \mu_i \Delta t + o(\Delta t). \quad (8)$$

Furthermore, the probability of two or more events (arrivals or departures) in an interval of length Δt is equal to $o(\Delta t)$.

Since, by independence, $\min\{A, S_i\} \sim \text{Exp}(\lambda + \mu_i)$, the random number $X(t_0 + \Delta t)$ of customers in the system at time $t_0 + \Delta t$ will be

$$\begin{cases} k+1 & \text{with probability } \lambda \Delta t + o(\Delta t), \\ k-1 & \text{with probability } \mu_i \Delta t + o(\Delta t), \\ k & \text{with probability } 1 - (\lambda + \mu_i) \Delta t + o(\Delta t). \end{cases} \quad (9)$$

Next, with the help of Bellman's principle of optimality, we may write that

$$\begin{aligned} F(k, t_0) &= \inf_{\substack{\mu(t) \\ t \in [t_0, T(k))}} \left\{ E \left[\int_{t_0}^{t_0 + \Delta t} \{q_0 \mu^2(t) + \lambda\} dt \right. \right. \\ &\quad \left. \left. + \int_{t_0 + \Delta t}^{T(k)} \{q_0 \mu^2(t) + \lambda\} dt \right] \right\} \\ &= \inf_{\substack{\mu(t) \\ t \in [t_0, t_0 + \Delta t]}} \left\{ [q_0 \mu^2(t_0) + \lambda] \Delta t + E[F(X(t_0 + \Delta t), t_0 + \Delta t)] \right\} \\ &\quad + o(\Delta t) \\ &= \inf_{\substack{\mu(t) \\ t \in [t_0, t_0 + \Delta t]}} \left\{ [q_0 \mu^2(t_0) + \lambda] \Delta t + F(k+1, t_0 + \Delta t) \lambda \Delta t \right. \\ &\quad + F(k-1, t_0 + \Delta t) \mu(t_0) \Delta t \\ &\quad \left. + F(k, t_0 + \Delta t) (1 - [\lambda + \mu(t_0)] \Delta t) \right\} \\ &\quad + o(\Delta t). \end{aligned} \quad (10)$$

We then obtain, through writing $\mu_0 := \mu(t_0)$,

$$\begin{aligned}
 F(k, t_0) - F(k, t_0 + \Delta t) = \inf_{\mu_0} \Big\{ & (q_0 \mu_0^2 + \lambda) \Delta t \\
 & + F(k+1, t_0 + \Delta t) \lambda \Delta t \\
 & + F(k-1, t_0 + \Delta t) \mu_0 \Delta t \\
 & - F(k, t_0 + \Delta t) (\lambda + \mu_0) \Delta t \Big\} \\
 & + o(\Delta t).
 \end{aligned} \tag{11}$$

Furthermore, using Taylor's formula, we can write that

$$F(\cdot, t_0 + \Delta t) = F(\cdot, t_0) + \Delta t F'(\cdot, t_0) + o(\Delta t). \tag{12}$$

It then follows that

$$F(k+1, t_0 + \Delta t) \lambda \Delta t = F(k+1, t_0) \lambda \Delta t + o(\Delta t). \tag{13}$$

Hence, through also applying Taylor's formula to the terms $F(k-1, t_0 + \Delta t) \mu_0 \Delta t$ and $F(k, t_0 + \Delta t) (\lambda + \mu_0) \Delta t$, we obtain that

$$\begin{aligned}
 F(k, t_0) - F(k, t_0 + \Delta t) = \inf_{\mu_0} \Big\{ & (q_0 \mu_0^2 + \lambda) \Delta t \\
 & + F(k+1, t_0) \lambda \Delta t \\
 & + F(k-1, t_0) \mu_0 \Delta t \\
 & - F(k, t_0) (\lambda + \mu_0) \Delta t \Big\} \\
 & + o(\Delta t).
 \end{aligned} \tag{14}$$

Finally, via dividing both sides of the previous equation by Δt and letting Δt decrease to zero, we obtain the following proposition:

Proposition 1. *If $S_i \sim \text{Exp}(\mu_i)$, for $i = 1, 2$, the value function $F(k, t_0)$ satisfies the following dynamic programming equation (DPE):*

$$\begin{aligned}
 -\frac{\partial}{\partial t_0} F(k, t_0) = \inf_{\mu_0} \Big\{ & q_0 \mu_0^2 + \lambda + F(k+1, t_0) \lambda + F(k-1, t_0) \mu_0 \\
 & - F(k, t_0) (\lambda + \mu_0) \Big\}.
 \end{aligned} \tag{15}$$

Moreover, we have the boundary conditions in Equation (5).

Remark 4. Equation (15) is valid for any $k \in \{2, 3, \dots\}$ if the capacity of the system is infinite. In the case when $c_p < \infty$, Equation (9) for $k = c_p$ reduces to

$$\begin{cases} k-1 & \text{with probability } \mu_i \Delta t + o(\Delta t), \\ k & \text{with probability } 1 - \mu_i \Delta t + o(\Delta t). \end{cases} \tag{16}$$

It follows that the dynamic programming equation becomes

$$-\frac{\partial}{\partial t_0} F(k, t_0) = \inf_{\mu_0} \Big\{ q_0 \mu_0^2 + \lambda + F(k-1, t_0) \mu_0 - F(k, t_0) \mu_0 \Big\} \tag{17}$$

for $k = c_p$. We could also assume, for instance, that if a customer arrives while the system is already full, then we stop the control problem and we set $F(c_p + 1, t_0) = F_2 \geq 0$. We would then have

$$-\frac{\partial}{\partial t_0} F(k, t_0) = \inf_{\mu_0} \left\{ q_0 \mu_0^2 + \lambda + F_2 \lambda + F(k-1, t_0) \mu_0 - F(k, t_0) (\lambda + \mu_0) \right\} \quad (18)$$

for $k = c_p$. In this paper, we assume that Equation (17) is the one that is valid if $k = c_p$.

Next, since $\mu_0 = \mu_1$ or μ_2 , Equation (15) can be rewritten as follows:

$$-\frac{\partial}{\partial t_0} F(k, t_0) = \min_{\mu_0 \in \{\mu_1, \mu_2\}} \left\{ q_0 \mu_0^2 + \lambda + F(k+1, t_0) \lambda + F(k-1, t_0) \mu_0 - F(k, t_0) (\lambda + \mu_0) \right\}. \quad (19)$$

To determine the value of $F(k, t_0)$, we can first consider the differential equation deduced from Equation (19) if the optimizer chooses $\mu_0 = \mu_i$ for any value of k and any t_0 , namely

$$-F'_i(k, t_0) = q_0 \mu_i^2 + \lambda + F_i(k+1, t_0) \lambda + F_i(k-1, t_0) \mu_i - F_i(k, t_0) (\lambda + \mu_i), \quad (20)$$

which is subject to the conditions in Equation (5). To obtain $F_i(k, t_0)$, we must, in general, solve a system of first-order linear differential equations.

Once the functions $F_i(k, t_0)$ have been computed for $i = 1, 2$ and for any k , we define it as follows:

$$\begin{aligned} G_i(k, t_0) &= q_0 \mu_i^2 + \lambda + \min\{F_1(k+1, t_0), F_2(k+1, t_0)\} \lambda \\ &\quad + \min\{F_1(k-1, t_0), F_2(k-1, t_0)\} \mu_i \\ &\quad - \min\{F_1(k, t_0), F_2(k, t_0)\} (\lambda + \mu_i) \end{aligned} \quad (21)$$

for $i = 1, 2$. The function $G_i(k, t_0)$ is the right-hand side of Equation (15) if $\mu_0 = \mu_i$. We can determine the optimal control by comparing the two functions for any $t_0 < t_1$.

Finally, to obtain the value function $F(k, t_0)$, we can solve the equation

$$-F'(k, t_0) = G_i(k, t_0) \quad (22)$$

in any interval where the optimal control is $\mu_0 = \mu_i$, and we can make use of the boundary condition $F(k, t_1) = F_1$, as well as the continuity of the function $F(k, t_0)$.

Remark 5. It is important to note that dynamic programming only works for the M/M/1 model (apart from a few special cases considered in other recent papers by the first author). Indeed, the results given in Equations (7) and (8) do not hold for general distributions. If $P[A \leq \Delta t]$ or $P[S_i \leq \Delta t]$ is not proportional to Δt , there will be a problem when we divide both sides of Equation (14) by Δt and we let Δt decrease to zero. Assuming that A and S_i are exponentially distributed is a simplifying assumption. Nevertheless, we believe that the M/M/1 model can serve as an approximate model in various applications.

We will now give an example that can be solved explicitly.

An Example

The simplest problem that can be considered is the one when $k = c_p = 2$. We then have the following equation (see Equation (17)):

$$-F'_i(2, t_0) = q_0 \mu_i^2 + \lambda - F_i(2, t_0) \mu_i. \quad (23)$$

The solution that satisfies the boundary condition $F_i(2, t_1) = F_1$ is

$$F_i(2, t_0) = \frac{q_0 \mu_i^2 + \lambda}{\mu_i} \left[1 - e^{-\mu_i(t_1 - t_0)} \right] + F_1 e^{-\mu_i(t_1 - t_0)}. \quad (24)$$

Let us take $\lambda = 1$, $\mu_1 = 2$, $\mu_2 = 3$, $q_0 = 1$, $F_1 = 20$ and $t_1 = 10$. We then have

$$F_1(2, t_0) = \frac{5}{2} + \frac{35}{2} e^{-2(10-t_0)} \quad (25)$$

and

$$F_2(2, t_0) = \frac{10}{3} + \frac{50}{3} e^{-3(10-t_0)} \quad (26)$$

for $t_0 \leq 10$. We can now compute the functions $G_1(2, t_0)$ and $G_2(2, t_0)$. These functions are displayed in Figures 1 and 2. The curves cross when $t_0 = s := 10 - \frac{1}{3} \ln(10)$. It follows that we then can write that the optimal control is

$$\mu_0^* = \begin{cases} 2 & \text{if } 0 \leq t_0 \leq s, \\ 3 & \text{if } s \leq t_0 < 10. \end{cases} \quad (27)$$

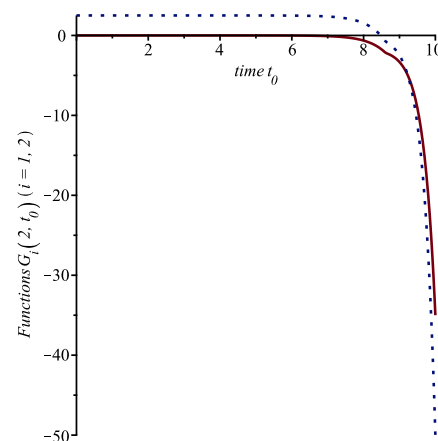


Figure 1. This figure illustrates the functions $G_1(2, t_0)$ and $G_2(2, t_0)$ on the interval $[0, 10]$. The solid line represents G_1 and the dotted line represents G_2 . These two functions were calculated on the basis of the parameters $\lambda = 1, \mu_1 = 2, \mu_2 = 3, q_0 = 1, F_1 = 20$ and $t_1 = 10$. The point of intersection of these two curves would be $s := 10 - 1/3 \ln(10)$, indicating the switching time from μ_1 to μ_2 for optimal service time control.

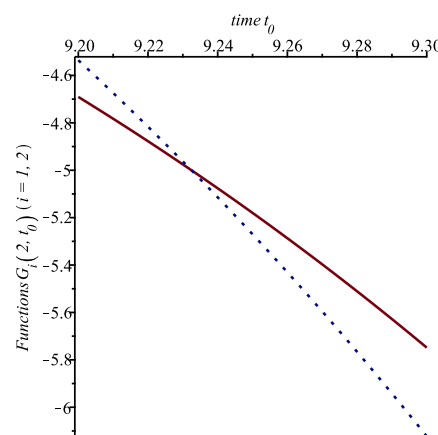


Figure 2. The functions $G_1(2, t_0)$ and $G_2(2, t_0)$ are zoomed in (on the critical interval $[9.2, 9.3]$). This figure illustrates the exact moment at which the intersection occurred and the control needed to be switched, providing further validation of the optimal change point calculated in Figure 1.

Finally, proceeding as explained above, we can calculate the value function $F(2, t_0)$, as shown in Figure 3.

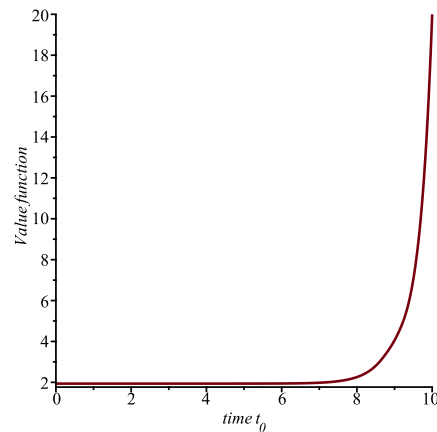


Figure 3. This figure illustrates the value function $F(2, t_0)$ in the interval $[0, 10]$. The parameters used are consistent with Figures 1 and 2 ($\lambda = 1, \mu_1 = 2, \mu_2 = 3, q_0 = 1, F_1 = 20$ and $t_1 = 10$). The value function $F(2, t_0)$ expresses the minimum cost expected given the optimal-control strategy obtained above.

In the next section, we will consider the case when the service time distributions are not necessarily exponential.

4. Optimal Control of the Modified M/G/1 Queue

Suppose that the optimizer must choose either the random service time S_1 or S_2 from the initial time t_0 to the final time $T(k)$ and that the function $K(\cdot)$ is identical to zero, so that the expected value of the cost function $J(k)$ defined in Equation (2) becomes (setting t_0 equal to zero)

$$E[J(k)] = E \left[\int_0^{T(k)} \left\{ \frac{q_0}{(E[S])^2} + \lambda \right\} dt \right] = \left(\frac{q_0}{(E[S])^2} + \lambda \right) E[T(k)], \quad (28)$$

where $S = S_1$ or S_2 . We will consider various particular cases below.

Case 1. Assume that $k = c_p = 2$. Since it is full, no new customers can enter the system between t_0 and $T(2)$. Moreover, we can write that $T(2)$ is simply equal to S . Thus, it follows that

$$E[J(2)] = \left(\frac{q_0}{(E[S])^2} + \lambda \right) E[S]. \quad (29)$$

In the special case when $S_i \sim \text{Exp}(\mu_i)$, for $i = 1, 2$, the above equation becomes

$$E[J(2)] = \left(q_0 \mu^2 + \lambda \right) \frac{1}{\mu}, \quad (30)$$

so that

$$F(2, t_0) = \min_{\mu \in \{\mu_1, \mu_2\}} \frac{q_0 \mu^2 + \lambda}{\mu}. \quad (31)$$

Suppose that $\mu_1 = 1, \mu_2 = 2$ and $\lambda = 1$. Then, we find that the optimal solution is $\mu^* = \mu_1$ if and only if $q_0 \geq 1/2$.

Remark 6. If we take the limit as t_1 tends to infinity in Equation (24), we obtain

$$F_i(2, t_0) = \frac{q_0 \mu_i^2 + \lambda}{\mu_i}, \quad (32)$$

which agrees with Equation (31) above.

Case 2. Let $k = 2$ and $c_p = 3$. We can write that

$$T(2) = \begin{cases} S_I & \text{if } S_I \leq A, \\ S_I + T(2) & \text{if } A < S_I, \end{cases} \quad (33)$$

where S_I is the time needed to serve the first customer, and $A \sim \text{Exp}(\lambda)$ is the time between successive customers, which is assumed to be independent of the service times. It then follows that

$$E[T(2)] = \frac{E[S_I]}{1 - P[A < S_I]}. \quad (34)$$

If $S_I \sim \text{Exp}(\mu)$, we have

$$1 - P[A < S_I] = P[S_I \leq A] = \frac{\mu}{\mu + \lambda}. \quad (35)$$

Thus,

$$E[T(2)] = \frac{\mu + \lambda}{\mu^2}. \quad (36)$$

To obtain the optimal solution, we must compare the value of $E[J(2)]$ deduced from Equation (28) and the above formula when $\mu = \mu_1$ with the corresponding one when $\mu = \mu_2$.

Remark 7. Let $Y_1, Y_2, \dots$ be independent random variables that have an $\text{Exp}(\mu)$ distribution, and let N be a geometric random variable with parameter

$$p := P[Y_1 < A] = \frac{\mu}{\mu + \lambda}. \quad (37)$$

We can then write that

$$E[T(2)] = E\left[\sum_{i=1}^N Y_i\right] = E[N]E[Y_i] = \frac{\mu + \lambda}{\mu} \frac{1}{\mu} = \frac{\mu + \lambda}{\mu^2}. \quad (38)$$

Hence, we retrieve the formula in Equation (36).

For the next cases, we will denote $T(k)$ by $T(k, c_p)$.

Case 3. When $k = 3$ and $c_p = 3$, no new customers can enter the system during the service time of the first customer. Hence, we can write that

$$T(3, 3) = S_I + T(2, 3) \implies E[T(3, 3)] = \frac{1}{\mu} + \frac{E[S_I]}{1 - P[A < S_I]}. \quad (39)$$

When $S_I \sim \text{Exp}(\mu)$, we obtain that

$$E[T(3, 3)] = \frac{1}{\mu} + \frac{\mu + \lambda}{\mu^2}. \quad (40)$$

Remark 8. In Section 2, to obtain the function $F_1(3, t_0)$, when $c_p = 3$, we must solve the system of first-order linear differential equations

$$-F_1'(3, t_0) = q_0 \mu_1^2 + \lambda + F_1(2, t_0) \mu_1 - F_1(3, t_0) \mu_1, \quad (41)$$

$$-F_1'(1, 2, t_0) = q_0 \mu_1^2 + \lambda + F_1(3, t_0) \lambda - F_1(2, t_0) (\lambda + \mu_1). \quad (42)$$

The solution to this system that satisfies the conditions in Equation (5) is easy to obtain. If we set $\lambda = 1$ and $\mu_1 = 3$, we find that the limit as t_1 tends to infinity of the solution is given by

$$F_1(2, t_0) = \frac{4}{9}(9q_0 + 1) \quad \text{and} \quad F_1(3, t_0) = \frac{7}{9}(9q_0 + 1). \quad (43)$$

Notice that $4/9 = E[T(2, 3)]$ and $7/9 = E[T(3, 3)]$ (see, respectively, Equations (36) and (40)).

Case 4. Finally, suppose that $c_p = 4$ and $k = 2$ or $k = 3$. We then define the event B_i as follows: there are exactly i customers that arrive during the service time S_I of the first customer, for $i = 0, 1, 2$, and we denote by A_I (respectively, A_{II}) the time needed for the first (and, respectively, second) new customer to arrive during the service time S_I . We can thus write that

$$T(3, 4) = \begin{cases} S_I + T(2, 4) & \text{if } B_0 \text{ occurs,} \\ S_I + T(3, 4) & \text{if } B_1 \text{ occurs.} \end{cases} \quad (44)$$

Similarly,

$$T(2, 4) = \begin{cases} S_I & \text{if } B_0 \text{ occurs,} \\ S_I + T(2, 4) & \text{if } B_1 \text{ occurs,} \\ S_I + T(3, 4) & \text{if } B_2 \text{ occurs.} \end{cases} \quad (45)$$

Hence, we have

$$E[T(3, 4)] = E[S_I] + E[T(2, 4)]P[S_I \leq A_I] + E[T(3, 4)]P[S_I > A_I] \quad (46)$$

and

$$\begin{aligned} E[T(2, 4)] &= E[S_I] + E[T(2, 4)]P[A_I < S_I, A_I + A_{II} \geq S_I] \\ &\quad + E[T(3, 4)]P[A_I + A_{II} < S_I]. \end{aligned} \quad (47)$$

To obtain $E[T(2, 4)]$ and $E[T(3, 4)]$, we can solve the above system of two linear equations. The various probabilities that appear in the system can at least be evaluated numerically in any particular case.

4.1. Particular Problem

We consider a particular problem in Case 4. Suppose that $S_1 \sim \text{Exp}(1)$ and $S_2 \sim G(2, 2)$, that is, S_2 has a gamma distribution with both parameters equal to 2, so that

$$f_{S_2}(t) = 4te^{-2t} \quad \text{for } t \geq 0. \quad (48)$$

Notice that $E[S_1] = E[S_2] = 1$. It then follows that the optimal control will depend only on the value of $E[T(k, 4)]$ for $k = 3$ and $k = 4$.

Let

$$p_{1i} := P[S_{Ii} \leq A_I], \quad (49)$$

$$p_{2i} := P[A_I + A_{II} < S_{Ii}] \quad (50)$$

and

$$p_{3i} := P[A_I < S_{Ii}, A_I + A_{II} \geq S_{Ii}], \quad (51)$$

where S_{Ii} is the random variable S_I if $S = S_i$ for $i = 1, 2$. We then find that

$$p_{11} = \frac{1}{\lambda + 1}, \quad p_{12} = \frac{4}{(\lambda + 2)^2}, \quad (52)$$

$$p_{21} = \frac{\lambda^2}{(\lambda + 1)^2}, \quad p_{22} = \frac{\lambda^2(\lambda + 6)}{(\lambda + 2)^3}, \quad (53)$$

$$p_{31} = \frac{\lambda}{(\lambda + 1)^2}, \quad p_{32} = \frac{8\lambda}{(\lambda + 2)^3}. \quad (54)$$

Notice that $\sum_{j=1}^3 p_{ji} = 1$ for $i = 1, 2$, as required.

Let $E[T_i(k, 4)]$ be the value of $E[T(k, 4)]$ if $S = S_i$, for $i = 1, 2$. To obtain $E[T_i(2, 4)]$ and $E[T_i(3, 4)]$, we must solve the following system of linear equations:

$$\begin{aligned} E[T_i(2, 4)] &= 1 + E[T_i(2, 4)] p_{3i} + E[T_i(3, 4)] p_{2i}, \\ E[T_i(3, 4)] &= 1 + E[T_i(2, 4)] p_{1i} + E[T(3, 4)] (1 - p_{1i}). \end{aligned} \quad (55)$$

We find that

$$E[T_1(2, 4)] = \lambda^2 + \lambda + 1, \quad E[T_2(2, 4)] = \frac{1}{16} \lambda^4 + \frac{1}{2} \lambda^3 + \lambda^2 + \lambda + 1, \quad (56)$$

$$E[T_1(3, 4)] = \lambda^2 + 2\lambda + 2, \quad E[T_2(3, 4)] = \frac{1}{16} \lambda^4 + \frac{1}{2} \lambda^3 + \frac{5}{4} \lambda^2 + 2\lambda + 2. \quad (57)$$

These functions are displayed in Figures 4 and 5, respectively.

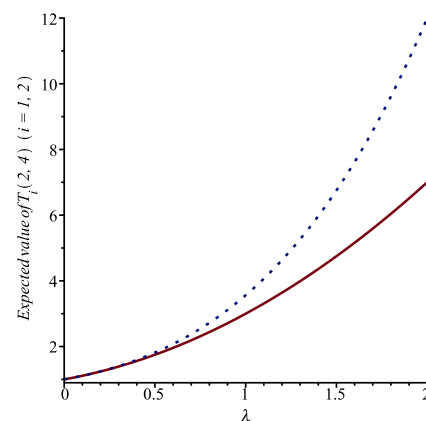


Figure 4. This figure shows the two expected time functions $E[T_1(2, 4)]$ and $E[T_2(2, 4)]$ for λ in the interval $[0, 2]$. The solid line represents $E[T_1(2, 4)]$ and the dotted line $E[T_2(2, 4)]$.

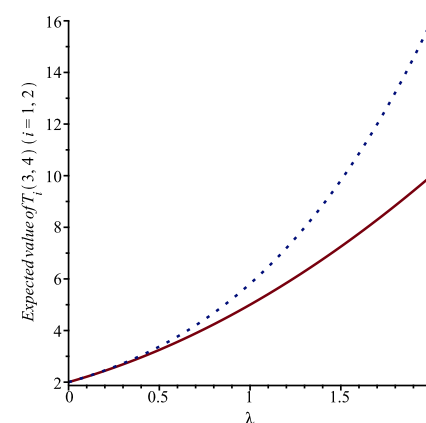


Figure 5. Like Figure 4, this figure shows the two expected time functions $E[T_1(3, 4)]$ and $E[T_2(3, 4)]$ over the same interval $[0, 2]$ for λ . The solid line represents $E[T_1(3, 4)]$ and the dotted line represents $E[T_2(3, 4)]$. These results highlight the distinction between queue management strategies in the case of variable service time distributions.

We can see that $E[T_1(2, 4)]$ and $E[T_2(2, 4)]$ are almost equal when λ is very small. However, as λ increases, $E[T_1(2, 4)]$ is much smaller than $E[T_2(2, 4)]$. In actuality, we can show that $E[T_1(2, 4)] < E[T_2(2, 4)]$ for any $\lambda > 0$. Hence, the optimal solution is always to choose $S = S_1$. The same conclusions apply to $E[T_1(3, 4)]$ and $E[T_2(3, 4)]$.

4.2. A Practical Application

We consider the problem of selecting an optimal-control policy for aircraft maintenance. In practice, maintenance programs must be optimized to minimize downtime and guarantee the expected level of safety, which are two fundamental elements in the aerospace industry. In our problem formulation, we must balance maintenance costs and the need to return the aircraft to service as soon as possible.

Suppose that S_1 is a Rayleigh random variable, so that

$$f_{S_1}(s) = \frac{s}{\sigma^2} \exp\left\{-\frac{s^2}{2\sigma^2}\right\} \quad \text{for } s \geq 0, \quad (58)$$

where σ is a positive constant and that

$$f_{S_2}(s) = \theta e^{-\theta s} \quad \text{for } s \geq 0. \quad (59)$$

That is, S_2 is an exponential random variable with parameter θ . The Rayleigh distribution, with its longer tail, should be more appropriate in the case of complex maintenance activities, while the exponential distribution could be used for customary maintenance activities.

The expected costs $E[J(3,4)]$ and $E[J(2,4)]$ for performing the maintenance tasks under each of the two different service time assumptions for a variable aircraft arrival rate for service were calculated. These costs include costs related to the speed of task execution (influenced by the service time distribution) and costs related to aircraft downtime (influenced by λ).

In Figure 6, we present the expected costs $E[J(3,4)]$ and $E[J(2,4)]$ when $\sigma = 3$, $\theta = 1$, $q_0 = 50$ and $\lambda \in [0.1, 5]$. Looking at the figure, we can conclude that, for routine inspections and predictable maintenance activities, the use of exponential service time distribution would be more cost effective, especially when arrival rates are high and rapid changes are required.

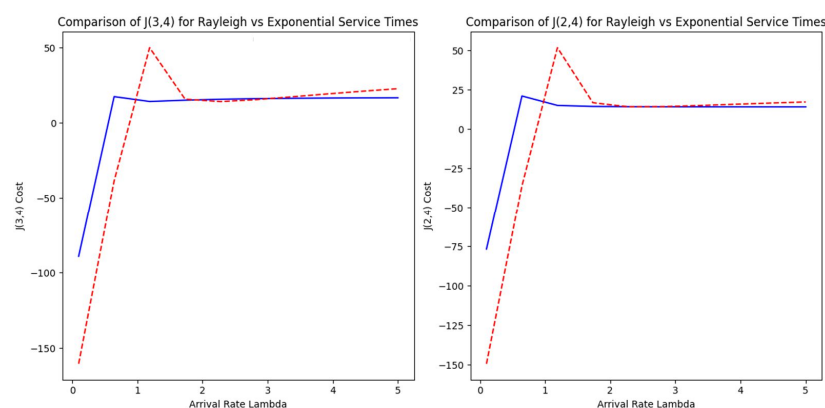


Figure 6. Expected cost $E[J(3,4)]$ (left panel) and $E[J(2,4)]$ (right panel) when S_1 has a Rayleigh distribution with parameter $\sigma = 3$ (full line) and S_2 has an exponential distribution with parameter $\theta = 1$ (dotted line), $q_0 = 50$ and $\lambda \in [0.1, 5]$.

For more complex activities that vary significantly over time (potentially due to unexpected complexities during the maintenance tasks), a more realistic modeling policy is offered by the Rayleigh distribution, although it may induce a higher cost. Such information would be essential for efficient planning and resource allocation, particularly in ensuring the availability of adequate maintenance personnel and spare parts.

Maintenance centers and airlines may decide to modify the planning of their maintenance decisions according to the historical frequency of maintenance needs (reflected by λ) and the complexity of tasks (complex or normal). For example, during busy seasons, it may be more cost effective to give priority to minimizing downtime (by choosing the exponential distribution).

5. Conclusions

We treated an optimal control problem for a queueing model with only one server. This queueing model was a modification of the classic $M/G/1$ model, in which the server can now choose between two distributions for the service time. The aim was to reduce the waiting queue to zero as soon as possible, while taking the quadratic control costs into account. The final time in the problem is the first time nobody is waiting for service in the system.

In Section 3, we assumed that the service time was exponentially distributed, with parameter μ_1 or μ_2 . We could then make use of dynamic programming to derive the equation satisfied by the value function, which gives the expected cost if the optimizer uses the optimal control from the initial time to the random final time.

Next, in Section 4 we generalized the problem to the case when the service time is any positive random variable. In general, it is then not possible to appeal to dynamic programming. We considered various particular cases when the capacity of the system is finite. We could in theory try to solve any particular problem. However, if the capacity of the system is large, the technique used to find the optimal solution will become rather intricate.

It is of course possible to consider other modifications of the basic $M/G/1$ queue. We could also assume that there is more than one server. The aim could be to use some control in order to increase the rate at which customers enter the system. If all the random variables in the model are exponentially distributed, we could try to utilize dynamic programming principles to obtain the optimal solution.

Author Contributions: Conceptualization, M.L.; methodology, M.L. and R.Y.; formal analysis, M.L.; writing—original draft, M.L.; writing—review & editing, R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Natural Sciences and Engineering Research Council of Canada

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: The authors wish to thank the anonymous reviewers of this paper for their constructive comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hosseinzadeh Lotfi, F.; Allahviranloo, T.; Pedrycz, W.; Shahriari, M.; Sharafi, H.; Razipour GhalehJough, S. Foundations of Decision. In *Fuzzy Decision Analysis: Multi Attribute Decision Making Approach*; Springer International Publishing: Cham, Switzerland, 2023; Volume 1121, pp. 1–56. [\[CrossRef\]](#)
2. Li, X.; Bi, S.; Zheng Y.; Wang, H. Energy-efficient online data sensing and processing in wireless powered edge computing systems. *IEEE Trans. Commun.* **2022**, *70*, 5612–5628. [\[CrossRef\]](#)
3. Aziati, A.H.N.; Hamdan, N.S.B. Application of queueing theory model and simulation to patient flow at the outpatient department. In Proceedings of the International Conference on Industrial Engineering and Operations Management, Bandung, Indonesia, 6–8 March 2018; pp. 3016–3028.
4. De, U.K.; Rajbongshi, G. Statistical application for the analysis of traffic congestion and its impact in a Hill city. *Int. J. Stat. Sci.* **2020**, *20*, 19–48.
5. Requia, W.J.; Higgins, C.D.; Adams, M.D.; Mohamed, M.; Koutrakis, P. The health impacts of weekday traffic: A health risk assessment of PM2.5 emissions during congested periods. *Environ. Int.* **2018**, *111*, 164–176. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Lefebvre, M.; Yaghoubi, R. Optimal control of a queueing system. 2024, *Submitted for publication*.

7. Laxmi, P.V.; Jyothisna, K. Optimization of service rate in a discrete-time impatient customer queue using particle swarm optimization. In *Distributed Computing and Internet Technology, Proceedings of the International Conference on Distributed Computing and Internet Technology, Bhubaneswar, India, 15–18 January 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 38–42. [\[CrossRef\]](#)
8. Tian, R.; Su, S.; Zhang, Z.G. Equilibrium and social optimality in queues with service rate and customers' joining decisions. *Qual. Technol. Quant. Manag.* **2024**, *21*, 1–34. [\[CrossRef\]](#)
9. Singh, S.K.; Cruz, G.M.B.; Cruz, F.R.B. Change point estimation of service rate in $M/M/1/m$ queues: A Bayesian approach. *Appl. Math. Comput.* **2024**, *465*, 128423. [\[CrossRef\]](#)
10. Laxmi, P.V.; Bhavani, E.G.; Jyothisna, K. Analysis of Markovian queueing system with second optional service operating under the triadic policy. *OPSEARCH* **2023**, *60*, 256–275. [\[CrossRef\]](#)
11. Wu, C.-H.; Yang, D.-Y.; Yong, C.-R. Performance evaluation and bi-objective optimization for F -policy queue with alternating service rates. *J. Ind. Manag. Optim.* **2023**, *19*, 3819–3839. [\[CrossRef\]](#)
12. Chen, G.; Liu, Z.; Xia, L. Event-based optimization of service rate control in retrial queues. *J. Oper. Res. Soc.* **2023**, *74*, 979–991. [\[CrossRef\]](#)
13. Dudin, A.; Dudin, S.; Dudina, O. Analysis of a queueing system with mixed service discipline. *Methodol. Comput. Appl. Prob.* **2023**, *25*, 19. [\[CrossRef\]](#)
14. Lakkumikanthan, I.; Balasubramanian, S. Optimal control of service rates of discrete-time (s, S) queueing—Inventory systems with finite buffer. In *Proceedings of the 5th International Conference on Problems of Cybernetics and Informatics, Baku, Azerbaijan, 28–30 August 2023*; pp. 1–4. [\[CrossRef\]](#)
15. Su, Y.; Li, J. Optimality of admission control in an $M/M/1/N$ queue with varying services. *Stoch. Models* **2021**, *37*, 317–334. [\[CrossRef\]](#)
16. Büke, B.; Qin, W. Many-server queues with random service rates in the Halfin-Whitt regime: A measure-valued process approach. *arXiv* **2019**. [\[CrossRef\]](#)
17. Dai, T.; Yu, T.; Zhao, X. Decision strategy of single server online customer service with variable service rates. In *Proceedings of the 2019 International Conference on Industrial Engineering and Systems Management, Shanghai, China, 25–27 September 2019*; pp. 1–6. [\[CrossRef\]](#)
18. Chen, S.; Xia, L. Optimal control of admission prices and service rates in open queueing networks. *IFAC-PapersOnLine* **2017**, *50*, 928–933. [\[CrossRef\]](#)
19. Rodrigues, J.; Prado, S.M.; Balakrishnan, N.; Louzada, F. Flexible $M/G/1$ queueing system with state dependent service rate. *Oper. Res. Lett.* **2016**, *44*, 383–389. [\[CrossRef\]](#)
20. Whittle, P. *Optimization over Time*; Wiley: Chichester, UK, 1982; Volume I.
21. Rishel, R. Controlled wear process: Modeling optimal control. *IEEE Trans. Automat. Control* **1991**, *36*, 1100–1102. [\[CrossRef\]](#)
22. Lefebvre, M. The homing problem for autoregressive processes. *IMA J. Math. Control Inform.* **2022**, *39*, 322–344. [\[CrossRef\]](#)
23. Lefebvre, M.; Pazhoheshfar, P. An optimal control problem for the maintenance of a machine. *Int. J. Syst. Sci.* **2022**, *53*, 3364–3373. [\[CrossRef\]](#)
24. Lefebvre, M. An optimal control problem for a modified $M/G/k$ queueing system. In *Proceedings of the Workshop on Intelligent Information Systems, Chişinău, Republic of Moldova, 19–21 October 2023*; pp. 141–148.
25. Lefebvre, M. Reducing the size of a waiting line optimally. *WSEAS Trans. Syst. Control.* **2023**, *18*, 342–345. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.