mathematics

Article

An M/G/1 Queue with Repeated Orbit While in Service

Gabi Hanukov [1,*], Yonit Barron [1] and Uri Yechiali [2]

[1] Department of Industrial Engineering & Management, Ariel University, Ariel 40700, Israel; ybarron@ariel.ac.il
[2] Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel; uriy@tauex.tau.ac.il
* Correspondence: gabih@ariel.ac.il; Tel.: +972-(0)504750747

Abstract: Orbit and retrial queues have been studied extensively in the literature. A key assumption in most of these works is that customers "go to orbit" when they are blocked upon arrival. However, real-life situations exist in which customers opt to go to orbit to efficiently use their orbit time rather than residing dormant at the service station while waiting for their service to be completed. This paper studies such a system, extending the scope of traditional orbit and retrial queues. We consider an M/G/1 queue where customers repeatedly go to orbit while their service remains in progress. That is, if a customer's service is not completed by within a specified "patience time", the customer goes to orbit for a random "orbit time". When the customer orbits, the server continues rendering her/his service. If, on return, the service is already completed, the customer leaves the system. Otherwise, s/he waits for another patience time. This policy is repeated until service completion. We analyze such an intricate system by applying the supplementary variable technique and using Laplace–Stieltjes transforms. Performance measures are derived, and a comparison analysis is provided between various service time distributions.

Keywords: orbit while in service; patience time; repeated trials; supplementary variable

MSC: 60K25; 90B22

Citation: Hanukov, G.; Barron, Y.; Yechiali, U. An M/G/1 Queue with Repeated Orbit While in Service. *Mathematics* **2024**, *12*, 3722. https://doi.org/10.3390/math12233722

Academic Editors: Oleg Tikhonenko and Marcin Ziółkowski

Received: 29 October 2024
Revised: 15 November 2024
Accepted: 19 November 2024
Published: 27 November 2024

1. Introduction

Orbit and retrial queues have been studied extensively in the queueing literature. The vast number of related works has been summarized in several publications, such as (i) the monograph by Falin and Tempelton [1], based on a careful analysis of several hundred publications; (ii) the comprehensive text by Artalejo and Gomez-Coral [2], which presents motivating examples in telephone and computer networks, while establishing a comparative analysis of retrial queues versus standard queues with waiting lines and queues with losses; (iii) a Special Issue (ANOR, Vol. 247) on Retrial Queues and Related Models edited by Gomez-Coral and Phung-Duc [3], comprising contributions by well-known experts in retrial and related queues; and many other references. A key assumption in most of these works is that customers who are blocked upon arrival immediately leave the system and "go to orbit" before their service even starts. After a random orbit time, they reattempt to gain admission to the service system. Many examples of retrial and orbit queues and their applications in daily life systems may be found in call centers, telecommunication systems, computer networks, and so-called "ticket queues" [4]. However, real-life situations exist in which customers may opt (or are obliged) to go to orbit while their service remains in progress. For example, consider a patient who passes a sequence of health examinations and related tests. It is unnecessary to wait for the results of one examination (e.g., blood test) before proceeding (i.e., "going to orbit") to the next stage. Such a procedure reduces the customer's overall treatment time. Another example is encountered in a supermarket delicatessen area where a customer may leave the area (i.e., go to orbit) and continue shopping while the order is being prepared. Similarly, a customer who brings her/his car to a garage and leaves the garage to perform some other errands before returning is a further example. Our model is also applicable to the construction area: when concrete is poured, the project is halted until the concrete is completely dry. This delay time is random, depending on external conditions. Thus, between the repeated tests, until the concrete is ready, the team is employed in other projects. Finally, we mention the field of aviation. Delaying flights to an unknown date forces the passenger to repeatedly check the status of their flight; in between checks, they utilize their time on other errands.

The above examples show that orbiting while in service may utilize customers' time and increase overall efficiency. Yet, it appears that the queueing literature has given minimal attention to these scenarios. Only recently, Hanukov and Yechiali [5] extended the scope of analysis of traditional orbit and retrial queues by studying an M/G/1 queueing system where impatient and time-use-oriented customers go to orbit while their service remains in progress to efficiently use their scarce time, rather than wait dormant at the service station until their service is completed. In that work, each customer served may exercise the "go to orbit" option no more than once during her/his service period. Nevertheless, customers may go to orbit and back multiple times while their service is in progress. Hanukov [6] analyzed this version for the M/M/1 queue with a limited number of customers in orbit by applying the matrix geometric technique. In the current paper, we further expand the scope of analysis of "orbit while in service" queues by studying an M/G/1 queueing system with an unlimited number of customers in orbit, where each customer may repeatedly go to orbit and back while her/his service is being processed.

Specifically, each customer whose service has begun (after reaching the head of the line) waits inside the system for a random amount of time (called "patience time") and then, if the service has not yet been completed, goes to orbit for a random amount of time (called "orbit time") before returning to the system to determine the status of her/his service. While the customer is in orbit, the server continues rendering service for that customer. On the customer's return, if the service has not yet been completed, s/he again waits inside the system for a "patience time," at the end of which, if the service is not yet completed, s/he takes another orbit time outside the system. The customer continues acting according to this policy until the service is completed, either when the customer is in the system or still orbiting. If the customer's service is completed while s/he is in orbit, the server proceeds to serve the next customers in the queue (if any). We aim to analyze this intricate system and derive its performance measures.

An important tool in our analysis is the supplementary variable technique (SVT), which is frequently used in the context of queueing systems analyses. In contrast to models having one (or more) bounded dimension, and thus allowing the use of matrix-geometric methods (see, e.g., [6]), the current model deals with a system having two unbounded dimensions of the state space, which leads to the use of SVT technique. In general, two kinds of supplementary variables are considered in the literature. One is the elapsed service (or arrival) time; the other is the remaining service (or arrival) time. The use of the SVT for the elapsed service time was first applied by Kosten [7] and Knessl et al. [8]. The method of using the remaining service time as the supplementary variable was first proposed by Henderson [9] and later used by Klimko and Neuts [10] and Minh [11]. The SVT has also been used in inventory models and reliability analyses [12]; see also Cohen [13] Ch. II.6.

In the current paper, applying the SVT enables the derivation of closed-form expressions for the probability generating function (PGF) and the mean number of customers in the system. Note that here we use the SVT for the elapsed service time, since using the SVT for the remaining service time mainly leads to recursive expressions (see, e.g., [12,14]).

Other contributions are as follows:

- We derive the Laplace–Stieltjes transforms (LSTs) for the following customer sojourn times: (a) V = the service-related sojourn time of a customer in the system calculated from the instant s/he starts service until departure, including time in orbit; and (b) H = the length of time a customer remains in orbit after her/his service has been completed.
- For an arbitrary customer, we obtain the mean number of orbits until departure.
- We provide a comparison study of various performance measures considering different service time distributions, including the family of gamma distributions (spanning the

range between the exponential and deterministic distributions), as well as the uniform distribution. The comparison study reveals the following qualitative conclusions:

(i). As customers spend more time in orbit and the service time variance decreases, a higher improvement is achieved in the system's performance.
(ii). If customers go more often to orbit (i.e., their patience time is shorter), or when they spend longer periods in orbit, their sojourn time inside the system is reduced, implying more efficient time use. However, their mean total time in the system (service plus orbit) is higher.
(iii). The mean total sojourn time in the system (inside and in orbit) is highest when the service time is deterministic, and the lowest values are obtained when the service time is exponential. Conversely, a lower mean sojourn time excluding orbit is achieved when the service time is deterministic, while the highest is when the service time is exponential.
(iv). The service time variance affects the mean number of orbits in opposite directions. In some cases, it is higher under the uniform distribution than under the exponential distribution, while in other cases, the opposite occurs.

The paper continues as follows: Section 2 presents a short literature review of related models. Section 3 describes the paper's model and depicts a flow scheme of the system. Applying the SVT, Section 4 derives the PGF and the mean number of customers in the system. Section 5 derives the LSTs of various sojourn times, and Section 6 presents a comparison study. Section 7 provides concluding remarks and managerial insights.

2. Related Literature

In addition to the manuscripts [1,2] and the special issue [3] cited in the introduction, many papers addressing orbit and retrial queues have appeared in the literature. For example, Kulkarni and Choi [15] consider a single server retrial queue where the server is subject to breakdowns and repairs. Two models are considered. In model I, the failed server cannot be occupied, and a customer whose service is interrupted either leaves the system or rejoins the retrial group. In model II, a customer whose service is interrupted by a failure remains with the server and restarts service when the repair process is completed. Recently, Kerner and Shmuel-Bittner [16] investigated strategic customer behavior in a hybrid M/M/1 queue with retrials. Both the Nash equilibrium and socially optimal retrial rates were studied. Fiems [17] studies the M/D/1 system with constant retrial times. The author derives explicit expressions for the distribution of the number of retrials and shows that it follows a geometric distribution. Dimitriou [18] considers a single-server retrial queue with event-dependent arrival rates. Performance measures are explicitly derived, and extensive numerical examples are provided, investigating the impact of event dependency. Jeganathan et al. [19] study a retrial queueing-inventory system where junior servers approach a single senior server on behalf of customers. Hu et al. [20] investigate customer retrials in call centers while considering service quality regarding the service speed. The authors reveal that high service quality and fast pick-up speed reduce retrials. It is revealed that, compared to business customers, private customers are more sensitive to quality but less sensitive to speed. Zhang and Wang [21] consider retrial queueing systems with boundedly rational customers. The authors find that the revenue-optimal price is no longer socially efficient generally but depends on the retrial rate. Furthermore, a threshold exists such that the socially optimal price exceeds the revenue-optimal one when the retrial rate is below this threshold; otherwise, the revenue-optimal price is greater. Xu et al. [22] analyze a retrial queueing system with priorities, where a new external arrival either expels the customer being served from the system and directly starts service or joins the retrial orbit. The server has multiple vacations whenever the orbit is empty of customers. Won et al. [23] consider an M/G/1 retrial queue with two types of calls: incoming calls made by regular customers and outgoing calls made by the server during its idle time. The service times of incoming and outgoing calls have non-identical distributions. The main interest of the paper is to analyze the waiting time distribution. Avrachenkov and Yechiali [24] investigate retrial networks and their application to internet data traffic, considering several benchmark retrial models. In all these works, customers go to orbit before joining the queue or before their service starts. Conversely, in our work, customers go to orbit *while their service remains in progress*.

Another stream of literature related to our work addresses so-called "impatient" customers who abandon the system if their waiting time before service initiation becomes too long. This topic has also been widely studied. Yechiali [25] investigates queues in which customers become impatient when the system is down and derives various quality-of-service measures: the mean sojourn time of a served customer, proportion of customers served, rate of lost customers due to disasters, and rate of abandonments due to impatience. Liu and Li [26] investigate a double-ended queue with first-come-first-match discipline under impatient customer behavior. The authors express the system as a level-dependent quasi-birth-and-death process with an infinite number of phases and provide a method to investigate the sojourn time of any arriving customer. Aalto [27] considers the optimal scheduling problem in a multi-server queue with impatient customers belonging to multiple classes. It is assumed that the scheduler cannot anticipate the expiration of the abandonment times but only knows their distributions and how long each customer has been in the system. Manitz and Piehl [28] investigate a multistage call center (comprising a front and a back office) and impatient customers. The main question is how many agents are necessary and how they should be allocated to maintain a service-level threshold and reduce the customer's expected waiting time. Lv et al. [29] analyze an M/M/c retrial queueing system with server working breakdown and impatient customers. Numerical experiments are conducted to assess the system's steady-state performance indicators. Altman and Yechiali [30,31] study cases where customers become impatient and leave the system only when the server is on vacation. Namely, an arriving customer who finds that the server is on vacation activates a random "impatience timer", such that if the system does not become available by the time the timer expires, the customer leaves the system. Inoue et al. [32] study a service system in which arriving customers are provided with information about the delay they will experience. Based on this information, they decide to wait for service or leave the system. Specifically, every customer has a patience threshold and balks if the observed delay exceeds the threshold. The main objective is to estimate the parameters of the customer patience-level distribution and the corresponding potential arrival rate, using knowledge of the actual queue-length process only. Bassamboo et al. [33] study scheduling multi-class impatient customers in queueing systems comprising parallel servers. The authors propose a multi-class time-in-queue policy that prioritizes across customer classes and within each class using a simple rule. For further discussion on customer impatience, see a recent survey paper [34] that develops a classification scheme of queueing models with customer impatience that includes vacation, feedback, priority service, and retention of reneging customers. It should be indicated that in all the above-mentioned "impatience" papers, the customers abandon the system permanently before their service has begun. Conversely, in our work, impatient customers do not leave the system but *repeatedly* go to orbit and back *during their service process* until it is completed.

It should be noted that models where customers "go to orbit" are inherently different from models where servers, not customers, go to "vacation", usually when the system becomes empty. Notable works on server vacations are [35–38]. For recent works dealing with vacations, among others, we mention, e.g., [39–41].

Another related set of models is queueing systems with feedback. In those models, if service is not successful, the customer repeats her/his service process. Recent related literature can be found in, e.g., [42–45]. In the latter papers, the customers may rejoin the system after their service has failed. In contrast, in our model, customers rejoin the system when the service is still in progress.

3. Model Description

Consider a single-server queueing system where customer arrival follows a Poisson process with rate λ. The service time is a random variable, B, with probability density function $f_B(t)$, cumulative distribution function $F_B(t)$, complementary cumulative distribution function $\overline{F}_B(t) = 1 - F_B(t)$, and LST $\equiv E[e^{-sB}] = \widetilde{B}(s)$. After a customer's service starts, s/he is willing to wait a random "patience time" T, exponentially distributed with parameter $\alpha(T \sim Exp(\alpha))$, for the service to be completed. If the service is completed before the patience time expires (i.e., $B \leq T$), the customer leaves the system on service completion. However, if the service has not been completed by time T (i.e., $T < B$), the customer goes to orbit for a random "orbit time" $X \sim Exp(\beta)$. While the customer is in orbit, the server continues rendering service for that customer. When the customer returns from orbit and finds that her/his service is already completed (i.e., $B < T + X$), s/he picks her/his order and leaves the system. Otherwise, s/he waits again in the system for another patience time $T_2 \sim T$. If the service has not been completed by that time $(T_2 < B - T - X)$, s/he leaves again for another orbit time, $X_2 \sim X$. This process is repeated until the customer's service is completed, whether s/he is inside the system or in orbit at the service completion instant. If a customer's service is completed while s/he is in orbit, the server continues serving the next customers (if any) in the queue. Figure 1 depicts the flow scheme of the system.

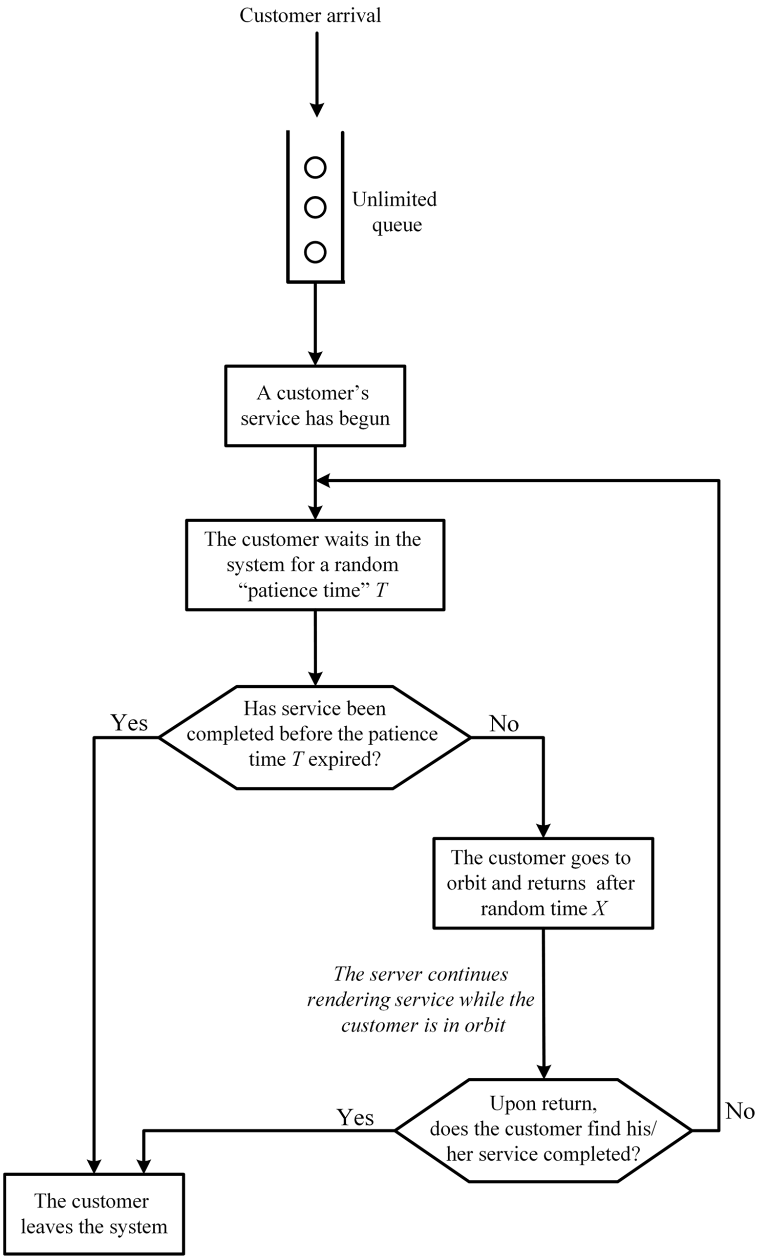

Figure 1. The flow scheme of the system.

4. Formulation and Analysis

Consider the system in a steady state. Let $L \in \{0,1,2,\ldots\}$ denote the number of customers in the system (not including those in orbit), and let $S \in \{1,2\}$ denote the status of the served customer as follows:

$$S = \begin{cases} 1 & \textit{the customer is in the system} \\ 2 & \textit{the customer is in orbit} \end{cases}$$

Let $U \in R^{+}$ denote the elapsed time since the start of service of the served customer (including orbit if applicable). Then, $\{L,S,U\}$ defines the system's state space. Let $p(n,m,u)$ denote the density function for the state $\{L=n,\ S=m,\ U=u\}$, $n=0,1,2,\ldots;\ m=1,2;\ u \geq 0$. Additionally, let $p(0)$ denote the probability of zero customers in the system and no orbiting customer in service.

We proceed to derive the mean number of customers in the system, not including those in orbit. For this purpose, we first define the following partial PGFs corresponding to the two possible values of $m=1,2$:

$$G_1(z,u) = \sum_{n=1}^{\infty} p(n,1,u)z^n,\ G_2(z,u) = \sum_{n=0}^{\infty} p(n,2,u)z^n$$

The partial PGFs, $G_1(z,u)$ and $G_2(z,u)$, are given in the following lemma.

Lemma 1. *$G_1(z,u)$ and $G_2(z,u)$ are given as a function of $p(0)$ as follows:*

$$G_1(z,u) = \frac{z\lambda(1-z)p(0)}{(\alpha+\beta)(\widetilde{B}(\lambda(1-z))-z)}\left(\beta e^{-\lambda(1-z)u}\overline{F}_B(u) + \alpha e^{-(\lambda(1-z)+\alpha+\beta)u}\overline{F}_B(u)\right) \quad (1)$$

$$G_2(z,u) = \frac{\alpha\lambda(1-z)p(0)}{(\alpha+\beta)(\widetilde{B}(\lambda(1-z))-z)}\left(e^{-\lambda(1-z)u}\overline{F}_B(u) - e^{-(\lambda(1-z)+\alpha+\beta)u}\overline{F}_B(u)\right) \quad (2)$$

Proof. We apply the SVT. Let $\mu(u)$ be the hazard rate function of the service time. Then, the density function $p(n,m,u)$ satisfies the following equations:

$$\begin{aligned} &p(n,1,u+h) = \lambda h p(n-1,1,u) + \beta h p(n-1,2,u) + [1-(\lambda+\mu(u)+\alpha)h]p(n,1,u) + o(h) \\ &n=1,2,3,\ldots,\ u>0, \text{where } p(0,1,u)=0, \text{and} \end{aligned} \quad (3)$$

$$\begin{aligned} &p(n,2,u+h) = \lambda h p(n-1,2,u) + \alpha h p(n+1,1,u) + [1-(\lambda+\mu(u)+\beta)h]p(n,2,u) + o(h) \\ &n=0,1,2,\ldots,\ u>0,\ \text{where } p(-1,2,u)=0, \end{aligned} \quad (4)$$

where $\lim\limits_{h\to 0}\frac{o(h)}{h}=0$.

The first term in Equations (3) and (4) refers to the arrival of a new customer. The second term of Equation (3) refers to the orbit termination, and the second term of Equation (4) refers to the patience time termination. The last term in Equations (3) and (4) refers to the corresponding complementary event.

Dividing Equations (3) and (4) by h, and letting $h \to 0$ leads to

$$\frac{d}{du}p(n,1,u) = \lambda p(n-1,1,u) + \beta p(n-1,2,u) - (\lambda+\mu(u)+\alpha)p(n,1,u),\ n=1,2,3,\ldots \quad (5)$$

$$\frac{d}{du}p(n,2,u) = \lambda p(n-1,2,u) + \alpha p(n+1,1,u) - (\lambda+\mu(u)+\beta)p(n,2,u),\ n=0,1,2,\ldots. \quad (6)$$

Multiplying Equations (5) and (6) by the corresponding z^n and summing over all n, we obtain the following set of differential equations for the two PGFs, $G_m(z,u)$, $m=1,2$:

$$\frac{d}{du}G_1(z,u) = [\lambda(z-1)-\mu(u)-\alpha]G_1(z,u) + \beta z G_2(z,u), \quad (7)$$

$$\frac{d}{du}G_2(z,u) = [\lambda(z-1) - \mu(u) - \beta]G_2(z,u) + \alpha z^{-1}G_1(z,u). \tag{8}$$

Using the Maple software, the solution of Equations (7) and (8) is given by

$$G_1(z,u) = C_1 e^{-\lambda(1-z)u - \int_0^u \mu(t)dt} + C_2 e^{-(\lambda(1-z)+\alpha+\beta)u - \int_0^u \mu(t)dt}, \tag{9}$$

$$G_2(z,u) = C_1\alpha(\beta z)^{-1} e^{-\lambda(1-z)u - \int_0^u \mu(t)dt} - C_2 z^{-1} e^{-(\lambda(1-z)+\alpha+\beta)u - \int_0^u \mu(t)dt} \tag{10}$$

Substituting the identity $\overline{F}_B(u) = e^{-\int_0^u \mu(t)dt}$ (see, e.g., Ross, page 389 [46]) into Equations (9) and (10) leads to

$$G_1(z,u) = C_1 e^{-\lambda(1-z)u}\overline{F}_B(u) + C_2 e^{-(\lambda(1-z)+\alpha+\beta)u}\overline{F}_B(u), \tag{11}$$

$$G_2(z,u) = C_1\alpha(\beta z)^{-1} e^{-\lambda(1-z)u}\overline{F}_B(u) - C_2 z^{-1} e^{-(\lambda(1-z)+\alpha+\beta)u}\overline{F}_B(u). \tag{12}$$

The constants C_1 and C_2 are derived by substituting the initial condition $u = 0$ into Equations (11) and (12). Setting $\overline{F}_B(0) = 1$ and $G_2(z,0) = 0$ leads to

$$G_1(z,0) = C_1 + C_2, \tag{13}$$

$$0 = C_1\alpha(\beta z)^{-1} - C_2 z^{-1}. \tag{14}$$

The solution of Equations (13) and (14) is

$$C_1 = \beta(\alpha+\beta)^{-1}G_1(z,0), \tag{15}$$

$$C_2 = \alpha(\alpha+\beta)^{-1}G_1(z,0). \tag{16}$$

Substituting Equations (15) and (16) into Equations (11) and (12), we obtain $G_1(z,u)$ and $G_2(z,u)$ as a function of $G_1(z,0)$:

$$G_1(z,u) = \frac{G_1(z,0)}{\alpha+\beta}\Big(\beta e^{\lambda(z-1)u}\overline{F}_B(u) + \alpha e^{(\lambda(z-1)-\alpha-\beta)u}\overline{F}_B(u)\Big), \tag{17}$$

$$G_2(z,u) = \frac{\alpha G_1(z,0)}{z(\alpha+\beta)}\Big(e^{\lambda(z-1)u}\overline{F}_B(u) - e^{(\lambda(z-1)-\alpha-\beta)u}\overline{F}_B(u)\Big). \tag{18}$$

To complete our analysis, we derive $G_1(z,0)$ as follows: Define

$$p(1,1,0) = \int_0^\infty [p(2,1,u) + p(1,2,u)]\mu(u)du + \lambda p(0), \tag{19}$$

$$p(n,1,0) = \int_0^\infty [p(n+1,1,u) + p(n,2,u))]\mu(u)du,\ n = 2,3,4,\ldots, \tag{20}$$

corresponding to the transition from state $(n+1,1,u)$ and state $(n,2,u)$ to state $(n,1,0)$ due to a service termination. The last term in Equation (19) expresses the event of a customer's arrival.

Multiplying Equations (19) and (20) by z^n and summing over all n, we obtain

$$zG_1(z,0) = \int_0^\infty (G_1(z,u) - zp(1,1,u) + zG_2(z,u) - zp(0,2,u))\mu(u)du + z^2\lambda p(0). \tag{21}$$

To derive the right-hand side of (21), we use the balance equation:

$$\lambda p(0) = \int_0^\infty [p(1,1,u) + p(0,2,u)]\mu(u))du. \quad (22)$$

Substituting Equation (22) into Equation (21) leads to

$$zG_1(z,0) = \int_0^\infty (G_1(z,u) + zG_2(z,u))\mu(u)du - z\lambda p(0) + z^2\lambda p(0). \quad (23)$$

Substituting $\overline{F}_B(u)\mu(u) = f_B(u)$, and Equations (17) and (18) into Equation (23), we obtain, after some algebra,

$$G_1(z,0) = \frac{z\lambda(1-z)p(0)}{(\widetilde{B}(\lambda(1-z)) - z)}. \quad (24)$$

The proof of Lemma 1 is completed by substituting Equation (24) into Equation (17) and into Equation (18). □

We further define $p(n,m) = P(L = n, S = m)$ and the two partial PGFs of the number of customers in the system:

$$G_1(z) = \sum_{n=1}^{\infty} p(n,1)z^n = \int_0^\infty G_1(z,u)du, \quad (25)$$

$$G_2(z) = \sum_{n=0}^{\infty} p(n,2)z^n = \int_0^\infty G_2(z,u)du. \quad (26)$$

Theorem 1. *The partial PGFs* $G_1(z)$ *and* $G_2(z)$ *are given by*

$$G_1(z) = \frac{z(1-\lambda E[B])}{(\alpha+\beta)(\widetilde{B}(\lambda(1-z)) - z)}\left[\beta(1-\widetilde{B}(\lambda(1-z))) + \frac{\lambda\alpha(1-z)(1-\widetilde{B}(\lambda(1-z)+\alpha+\beta))}{((\lambda(1-z)+\alpha+\beta))}\right], \quad (27)$$

$$G_2(z) = \frac{\alpha(1-\lambda E[B])}{(\alpha+\beta)(\widetilde{B}(\lambda(1-z)) - z)}\left[(1-\widetilde{B}(\lambda(1-z))) - \frac{\lambda(1-z)(1-\widetilde{B}(\lambda(1-z)+\alpha+\beta))}{(\lambda(1-z)+\alpha+\beta)}\right]. \quad (28)$$

Proof. For an arbitrary function $g(z)$, we use the identity

$$\int_0^\infty e^{-g(z)u}\overline{F}_B(u)du = \frac{1-\widetilde{B}(g(z))}{g(z)} \quad (29)$$

(see Ross, page 390 [46]). Then, substituting Equations (24)–(26) into Equations (17) and (18), and applying identity (29), the PGFs can be expressed as functions of $p(0)$ as follows:

$$G_1(z) = p(0)\left[\frac{z\beta(1-\widetilde{B}(\lambda(1-z)))}{(\alpha+\beta)(\widetilde{B}(\lambda(1-z)) - z)} + \frac{z\lambda\alpha(1-z)(1-\widetilde{B}(\lambda(1-z)+\alpha+\beta))}{(\alpha+\beta)(\widetilde{B}(\lambda(1-z)) - z)((\lambda(1-z)+\alpha+\beta))}\right], \quad (30)$$

$$G_2(z) = p(0)\left[\frac{\alpha(1-\widetilde{B}(\lambda(1-z)))}{(\alpha+\beta)(\widetilde{B}(\lambda(1-z)) - z)} - \frac{\lambda\alpha(1-z)(1-\widetilde{B}(\lambda(1-z)+\alpha+\beta))}{(\alpha+\beta)(\widetilde{B}(\lambda(1-z)) - z)((\lambda(1-z)+\alpha+\beta))}\right]. \quad (31)$$

Now, $p(0)$ is calculated by solving $p(0) + G_1(z = 1) + G_2(z = 1) = 1$ and using L'Hôpital's rule. As expected, $p(0)$ is given by

$$p(0) = 1 - \lambda E[B]. \tag{32}$$

Substituting Equation (32) into Equations (30) and (31) proves Theorem 1. □

Two points should be mentioned.

(i). The time that customers spend in orbit does not affect the utilization of the server, as expressed in Equation (32).

(ii). If the customer's patience time is unbounded, or if the orbit time tends to zero, we obtain the PGF of the number of customers in the classical M/G/1 queue, namely,

$$p(0) + G_1(z) + G_2(z) \underset{\alpha \to 0}{\to} \frac{(1-z)\widetilde{B}(\lambda(1-z))}{\widetilde{B}(\lambda(1-z)) - z}(1 - \lambda E[B]),$$

$$p(0) + G_1(z) + G_2(z) \underset{\beta \to \infty}{\to} \frac{(1-z)\widetilde{B}(\lambda(1-z))}{\widetilde{B}(\lambda(1-z)) - z}(1 - \lambda E[B]).$$

Finally, the mean number of customers in the system, $E[L]$, is given in the next theorem.

Theorem 2.

$$E[L] = \lambda E[B] + \frac{\lambda^2 E[B^2]}{2(1 - \lambda E[B])} - \frac{\alpha\lambda}{\alpha + \beta}\left(E[B] - \frac{1 - \widetilde{B}(\alpha + \beta)}{\alpha + \beta}\right). \tag{33}$$

Proof. Applying $E[L] = \frac{dG_1(z)}{dz}\Big|_{z=1} + \frac{dG_2(z)}{dz}\Big|_{z=1}$ and using L'Hôpital's rule leads to Equation (30). □

Note that Equation (33) can be written as

$$E[L] = E[L_{M/G/1}] - \frac{\alpha\lambda}{\alpha + \beta}\left(E[B] - \frac{1 - \widetilde{B}(\alpha + \beta)}{\alpha + \beta}\right),$$

where $E[L_{M/G/1}]$ is the mean number of customers in the system in the classical M/G/1 queue.

Stability condition. It is instructive to emphasize that the system's stability condition is identical to that of the classical M/G/1 queue, namely, $\lambda E[B] < 1$. This follows since service is not interrupted and the server continues rendering service when the customers go to orbit. Formally, the above condition is obtained as a result of $p(0) > 0$ (see, Equation (32)).

5. Customer's Sojourn Times and Number of Orbits

In this section, we derive LSTs of two important sojourn times and calculate the mean number of orbits of an arbitrary customer. Specifically, we focus on (i) a customer's sojourn time from the moment s/he starts service until her/his departure (including orbit time), (ii) the orbit time of a customer after her/his service is completed, and (iii) the number of orbit excursions made by an arbitrary customer.

To this end, we first define the following random variables.

Let $T(k) = \sum_{i=1}^{k} T_i$ be the sum of k independent patience times T_i, each distributed as T. Since $T \sim Exp(\alpha)$, $T(k)$ is Erlang distributed with k stages, that is, $T(k) \sim Erlang(k, \alpha)$.

Similarly, let $X(k) = \sum_{i=1}^{k} X_i$ be the sum of k independent orbit times X_i, each distributed as $X \sim Exp(\beta)$, implying that $X(k) \sim Erlang(k, \beta)$.

Define $Y(k) = T(k) + X(k)$. The probability distribution function of $Y(k)$ is given by

$$\begin{aligned} F_{Y(k)}(y) \equiv P(Y(k) \leq y) &= \int_{t=0}^{y} f_{T(k)}(t) \int_{x=0}^{y-t} f_{X(k)}(x)\,dx\,dt \\ &= \int_{t=0}^{y} \frac{\alpha^k t^{k-1} e^{-\alpha t}}{(k-1)!} \int_{x=0}^{y-t} \frac{\beta^k x^{k-1} e^{-\beta x}}{(k-1)!}\,dx\,dt. \end{aligned} \tag{34}$$

The density function of $Y(k)$ is

$$f_{Y(k)}(y) = \frac{dF_{Y(k)}(y)}{dy}. \tag{35}$$

For example, for $k = 1$, we have

$$F_{Y(1)}(y) = 1 + \frac{\beta e^{-\alpha y} - \alpha e^{-\beta y}}{\alpha - \beta} \text{ and } f_{Y(1)}(y) = \frac{\alpha\beta(e^{-\beta y} - e^{-\alpha y})}{\alpha - \beta}.$$

For $k = 2$, we obtain

$$F_{Y(2)}(y) = 1 + \frac{\beta^2(\beta - 3\alpha - \alpha(\alpha - \beta)y)e^{-\alpha y} - \alpha^2(\alpha - 3\beta - \beta(\beta - \alpha)y)e^{-\beta y}}{(\alpha - \beta)^3},$$

$$f_{Y(2)}(y) = \frac{\alpha^2\beta^2(((\beta - \alpha)y - 2)e^{-\alpha y} - ((\alpha - \beta)y - 2)e^{-\beta y})}{\alpha - \beta}.$$

Let V denote the sojourn time of a customer in the system from the moment s/he starts service until her/his departure, including the time in orbit.

Theorem 3. *The LST of V is given by*

$$\begin{aligned} \widetilde{V}(s) = &\int_{b=0}^{\infty} \int_{t=0}^{b} \int_{x=b-t}^{\infty} e^{-s(t+x)} f_B(b) f_T(t) f_X(x) dx\,dt\,db \\ &+ \sum_{k=2}^{\infty} \int_{b=0}^{\infty} \int_{y=0}^{b} \int_{t=0}^{b-y} \int_{x=b-y-t}^{\infty} e^{-s(y+t+x)} f_B(b) f_{Y(k-1)}(y) f_T(t) f_X(x) dx\,dt\,dy\,db \\ &+ \int_{b=0}^{\infty} \int_{t=b}^{\infty} e^{-sb} f_B(b) f_T(t) dt\,db + \sum_{k=1}^{\infty} \int_{b=0}^{\infty} \int_{y=0}^{b} \int_{t=b-y}^{\infty} e^{-sb} f_B(b) f_{Y(k)}(y) f_T(t) dt\,dy\,db. \end{aligned} \tag{36}$$

Proof. Two possible scenarios exist: (i) the customer goes to orbit k times, and the service is completed while s/he is in orbit. Under this scenario, the customer's sojourn time is $Y(k)$, the sum of k independent patience times plus k independent orbit times; (ii) the customer goes to orbit k times, and the service is completed while s/he is in the system. Under this scenario, the customer's sojourn time is equal to the service time, B. Considering these two scenarios, V is equal to

$$V = \begin{cases} Y(k) & if \;\; Y(k-1) + T_k < B < Y(k), \;\; k = 1, 2, 3, \ldots, \\ B & if \;\; Y(k) < B < Y(k) + T_{k+1}, \;\; k = 0, 1, 2, 3, \ldots. \end{cases} \tag{37}$$

Note that T_k refers to a customer's k^{th} patience time, where all T_k are independent and identically distributed (i.i.d.) as T.

Applying Equation (37), the LST of V is given by

$$\begin{aligned}\widetilde{V}(s) = E\Bigg[&\sum_{k=1}^{\infty} E[e^{-sY(k)}|Y(k-1)+T_k < B < Y(k)]P(Y(k-1)+T_k < B < Y(k)) \\ &+\sum_{k=0}^{\infty} E[e^{-sB}|Y(k) < B < Y(k)+T_{k+1}]P(Y(k) < B < Y(k)+T_{k+1})\Bigg].\end{aligned} \tag{38}$$

The first line of Equation (38) leads to the first two lines in Equation (36), and the second line of Equation (38) leads to the third line in Equation (36). This completes the proof. □

Now, a customer's mean total service time $E[V]$ is given by

$$E[V] = -\left.\frac{d\widetilde{V}(s)}{ds}\right|_{s=0}. \tag{39}$$

Next, let H denote the time a customer remains in orbit after her/his service has been completed.

Theorem 4. *The LST of H is given by*

$$\begin{aligned}\widetilde{H}(s) = &\int_{b=0}^{\infty}\int_{t=0}^{b}\int_{x=b-t}^{\infty} e^{-s(t+x-b)} f_B(b) f_T(t) f_X(x)\,dx\,dt\,db \\ &+\sum_{k=2}^{\infty}\int_{b=0}^{\infty}\int_{y=0}^{b}\int_{t=0}^{b-y}\int_{x=b-y-t}^{\infty} e^{-s(y+t+x-b)} f_B(b) f_{Y(k-1)}(y) f_T(t) f_X(x)\,dx\,dt\,dy\,db \\ &+1-\int_{b=0}^{\infty}\int_{t=0}^{b}\int_{x=b-t}^{\infty} f_B(b) f_T(t) f_X(x)\,dx\,dt\,db \\ &-\sum_{k=2}^{\infty}\int_{b=0}^{\infty}\int_{y=0}^{b}\int_{t=0}^{b-y}\int_{x=b-y-t}^{\infty} f_B(b) f_{Y(k-1)}(y) f_T(t) f_X(x)\,dx\,dt\,dy\,db.\end{aligned} \tag{40}$$

Proof. Two possible scenarios exist: (i) the customer goes to orbit k times, and her/his service is completed while s/he is in orbit. Under this scenario, the time a customer remains in orbit after her/his service has been completed is equal to $Y(k) - B$, that is, the sum of k independent patience times plus k independent orbit times minus the generic service time B. (ii) For other cases, the customer does not remain in orbit after service completion.

Considering these two scenarios, H is equal to

$$H = \begin{cases} Y(k) - B \;\; if \;\; Y(k-1)+T_k < B < Y(k), \;\; k = 1,2,3,\ldots, \\ 0 \qquad Otherwise. \end{cases} \tag{41}$$

Applying Equation (41), the LST of H is given by

$$\begin{aligned}\widetilde{H}(s) = &E[e^{-s(Y(k)-B)}|Y(k-1)+T_k < B < Y(k)]P(Y(k-1)+T_k < B < Y(k)) \\ &+1\cdot(1-P(Y(k-1)+T_k < B < Y(k))).\end{aligned} \tag{42}$$

Equation (42) leads to Equation (40), which proves Theorem 4. □

Thus, the mean remaining orbit time after service completion is given by

$$E[H] = -\left.\frac{d\widetilde{H}(s)}{ds}\right|_{s=0}. \tag{43}$$

Let O denote the number of orbit excursions made by an arbitrary customer before leaving the system.

Theorem 5. *The mean number of orbits until departure, $E[O]$, is calculated by*

$$E[O] = \sum_{k=1}^{\infty} kP(O = k) = \sum_{k=1}^{\infty} k[P(Y(k-1) + T_k < B < Y(k)) + P(Y(k) < B < Y(k) + T_{k+1})], \tag{44}$$

which is explicitly given by

$$\begin{aligned} E[O] = 1 \cdot & \int_{b=0}^{\infty} \int_{t=0}^{b} \int_{x=b-t}^{\infty} f_B(b) f_T(t) f_X(x) dx\, dt\, db \\ & + \sum_{k=2}^{\infty} k \int_{b=0}^{\infty} \int_{y=0}^{b} \int_{t=0}^{b-y} \int_{x=b-y-t}^{\infty} f_B(b) f_{Y(k-1)}(y) f_T(t) f_X(x) dx\, dt\, dy\, db \\ & + \sum_{k=1}^{\infty} k \int_{b=0}^{\infty} \int_{y=0}^{b} \int_{t=b-y}^{\infty} f_B(b) f_{Y(k)}(y) f_T(t) dt\, dy\, db. \end{aligned} \tag{45}$$

Proof. The above expression for $E[O]$ follows directly from Equation (37).

6. Impact of Different Service Time Distributions

In this section, we numerically investigate the impact of different service time distributions on the system's performance measures. Specifically, we focus on five measures: (1) mean number of customers in the system (Figures 2 and 3); (2) mean sojourn time excluding orbit time (Figures 4 and 5); (3) mean total customer sojourn time (including orbit; Figures 6 and 7); (4) mean remaining orbit time after service completion (Figures 8 and 9); and (5) mean number of orbits by a customer (Figures 10–12). Additional Figures S1–S5 are presented in a Supplementary Materials File.

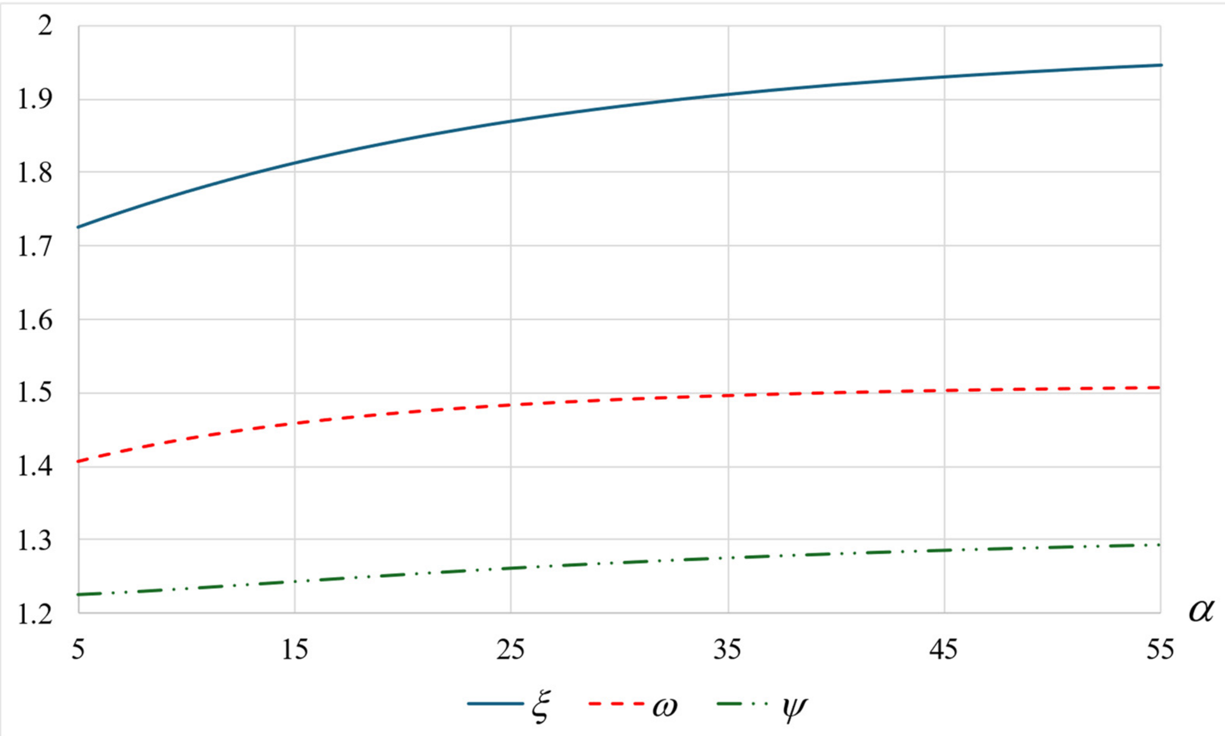

Figure 2. ξ, ω, and ψ as a function of α.

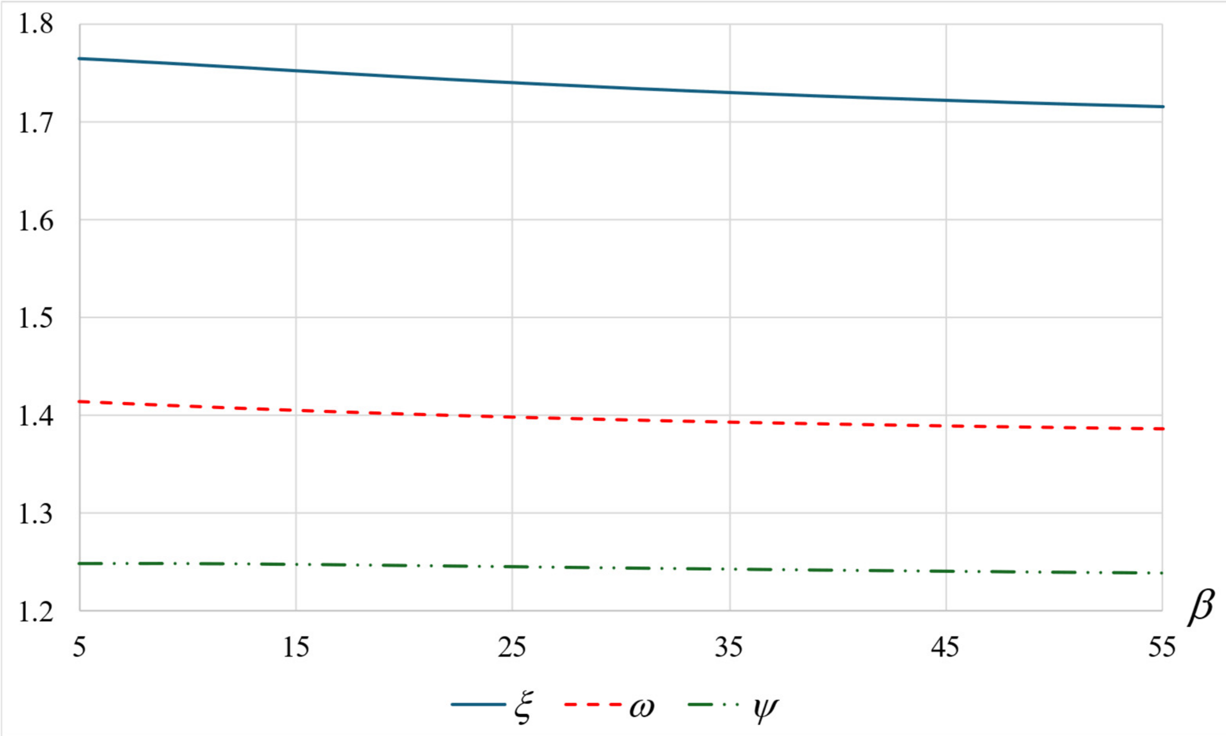

Figure 3. ξ, ω, and ψ as a function of β.

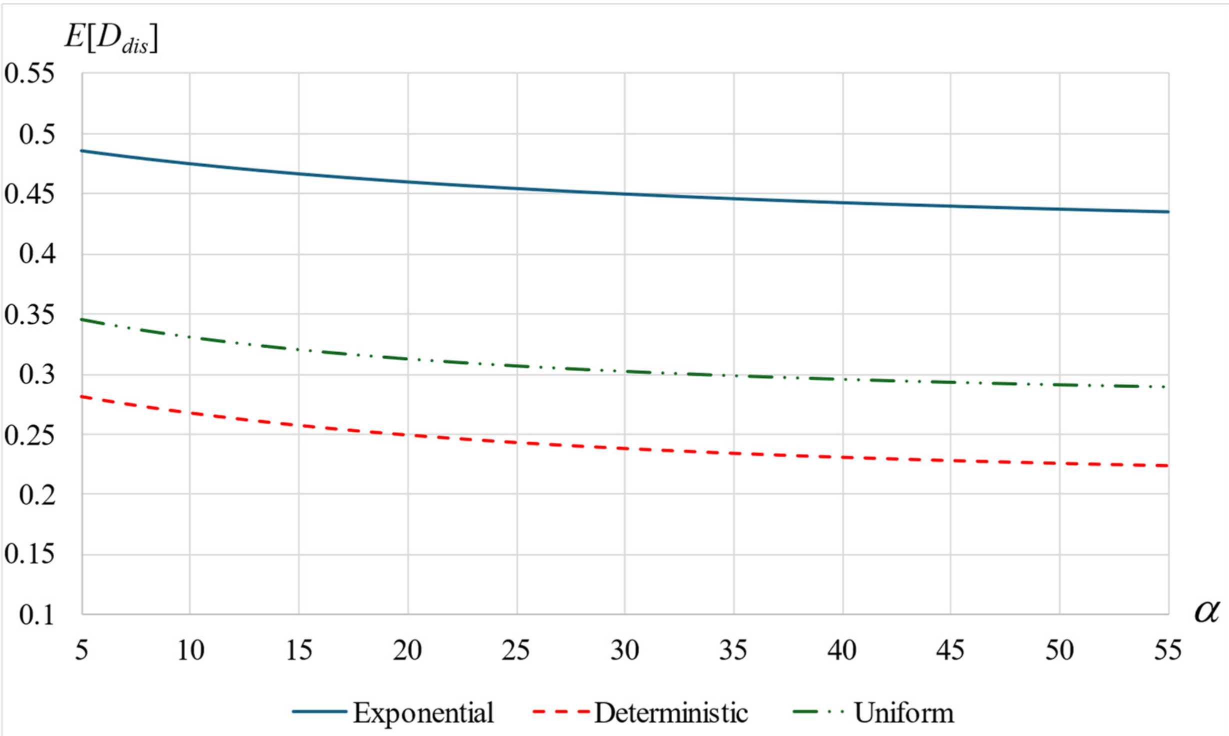

Figure 4. $E[D_{dis}]$ as a function of α.

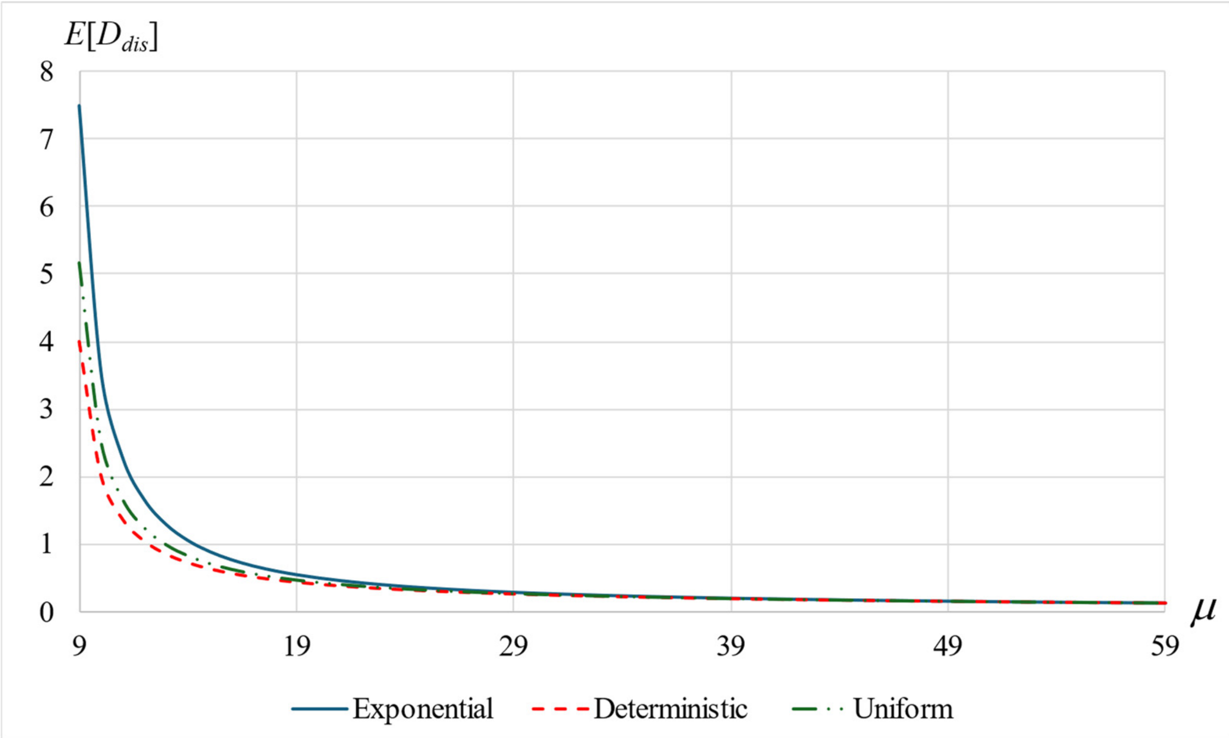

Figure 5. $E[D_{dis}]$ as a function of μ.

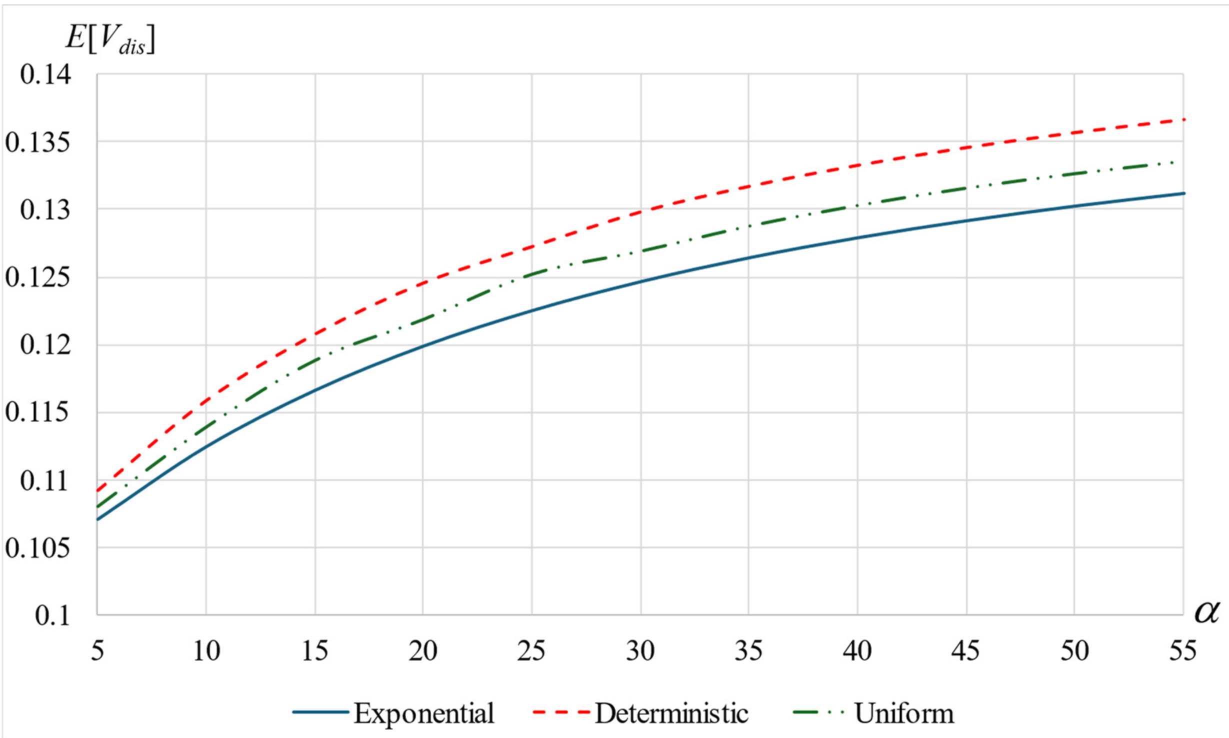

Figure 6. $E[V_{dis}]$ as a function of α.

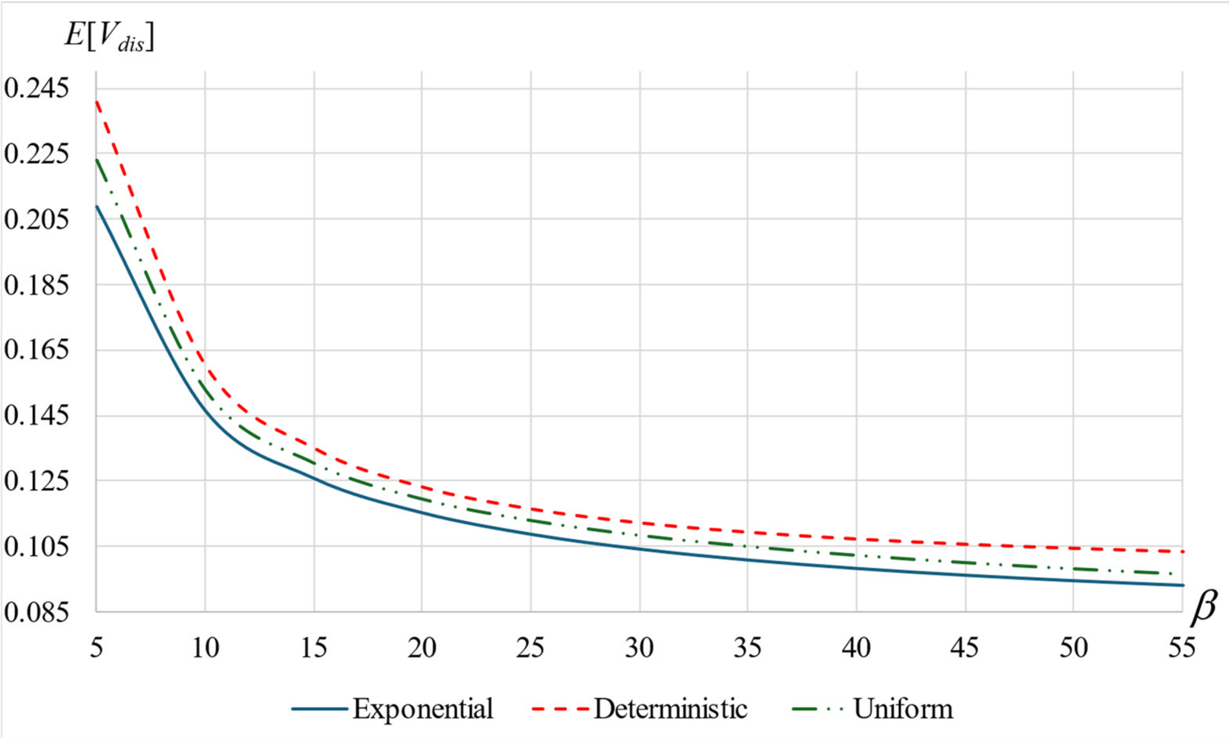

Figure 7. $E[V_{dis}]$ as a function of β.

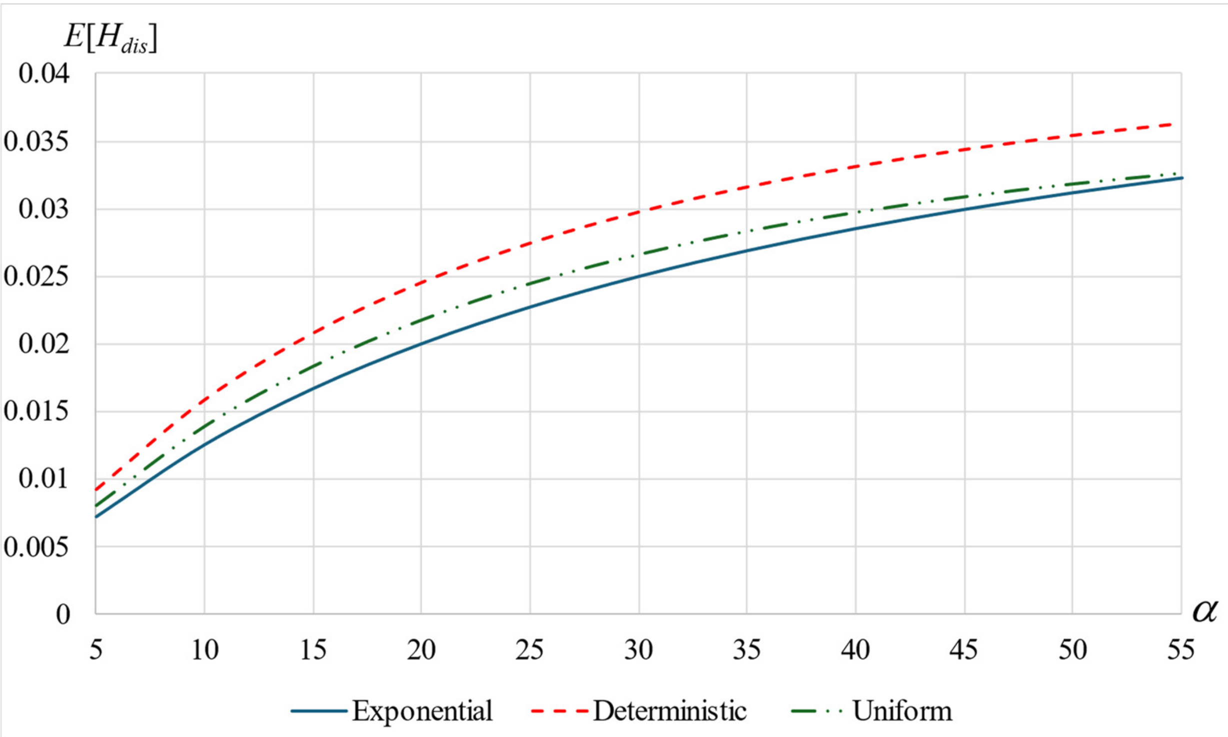

Figure 8. $E[H_{dis}]$ as a function of α.

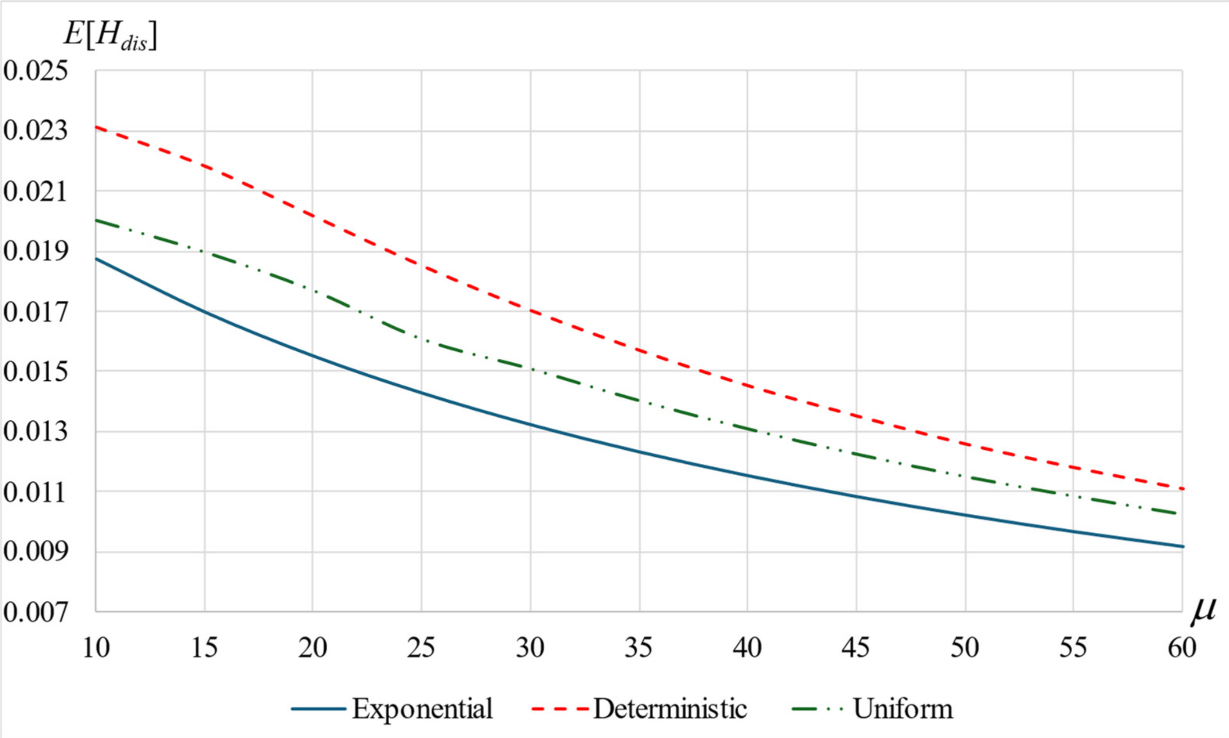

Figure 9. $E[H_{dis}]$ as a function of μ.

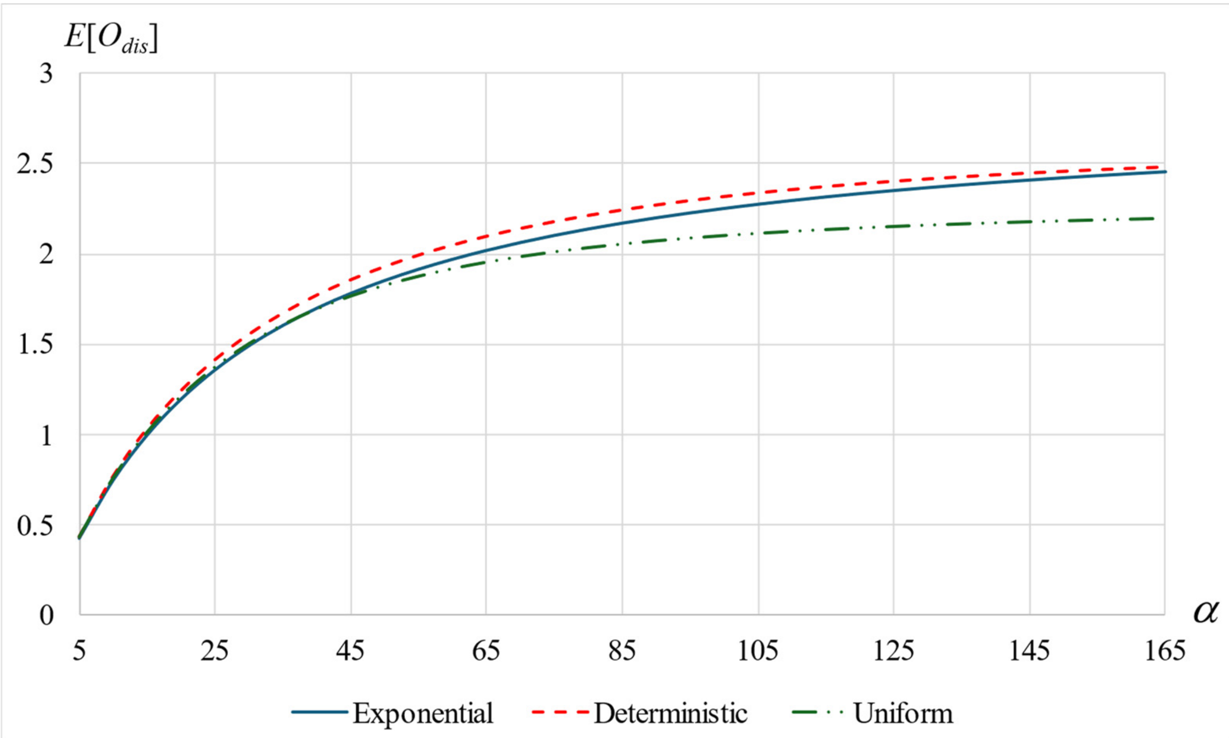

Figure 10. $E[O_{dis}]$ as a function of α.

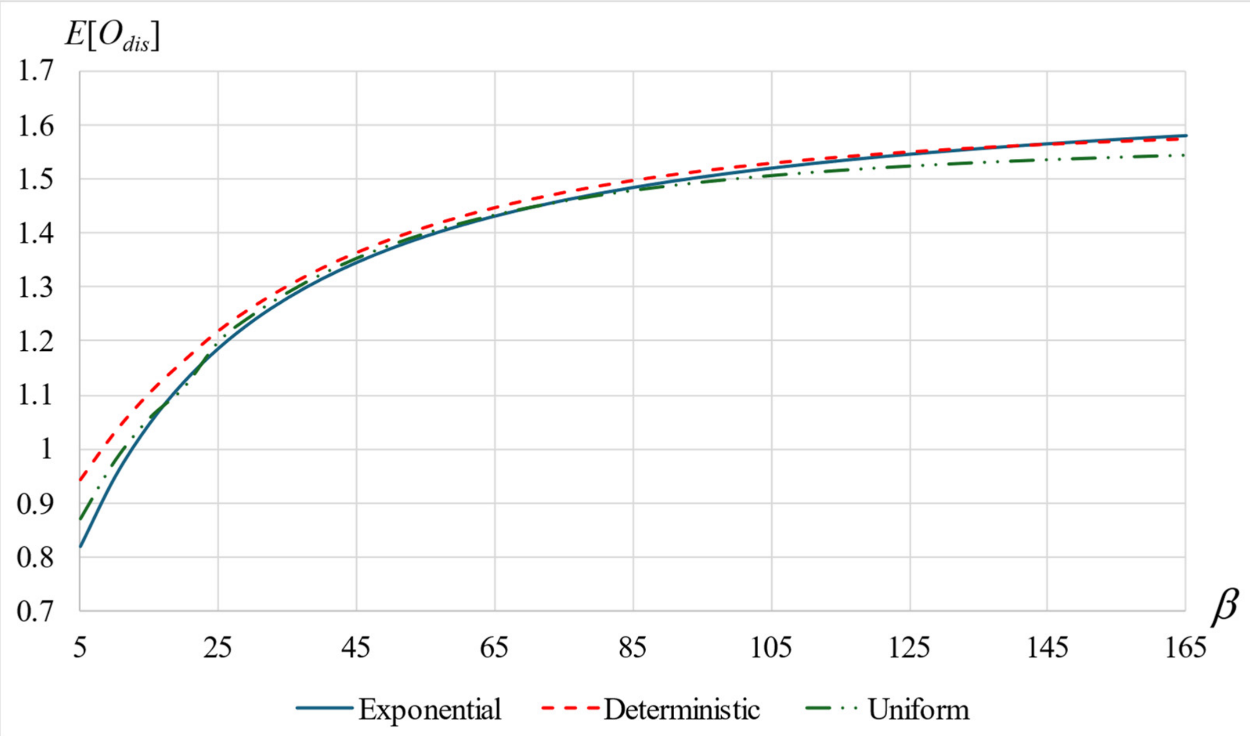

Figure 11. $E[O_{dis}]$ as a function of β.

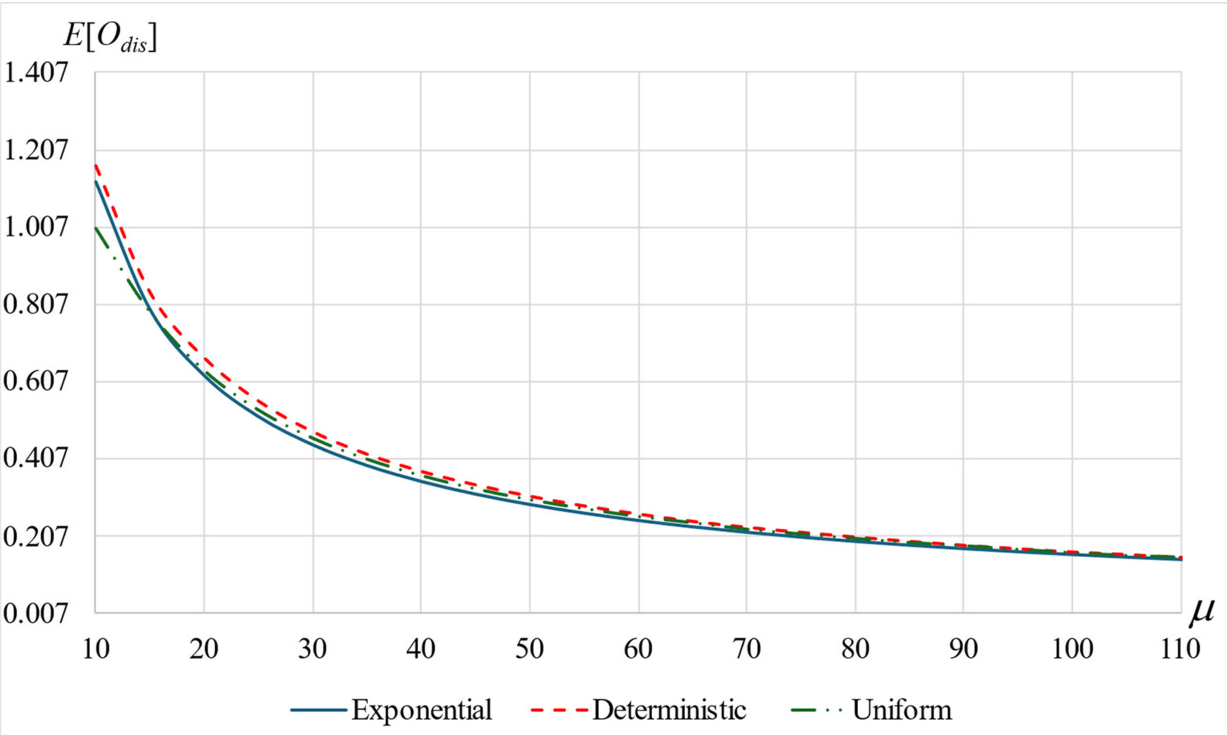

Figure 12. $E[O_{dis}]$ as a function of μ.

We consider the family of (a) gamma distribution functions, spanning the range between the exponential distribution having coefficient of variation $CV = 1$ and the deterministic distribution with $CV = 0$, and (b) the uniform distribution.

Specifically, we consider the following:

(i). The *gamma distribution* with scale γ and shape $\gamma\mu$. That is, $B \sim \Gamma(\gamma, \gamma\mu)$, with $E[B] = \frac{\gamma}{\gamma\mu} = \frac{1}{\mu}$, $E[B^2] = \frac{1+\gamma}{\gamma\mu^2}$, $V[B] = \frac{1}{\gamma\mu^2}$, LST $\widetilde{B}(s) = \left(\frac{\gamma\mu}{\gamma\mu+s}\right)^{\gamma}$, and $CV^2 = \frac{1}{\gamma}$.

(ii). The *exponential distribution*, $B \sim Exp(\mu)$, which is a special extreme case of the gamma distribution with $\gamma = 1$.

(iii). The *deterministic distribution* obtained from the gamma distribution by letting $\gamma \to \infty$.

(iv). The *uniform distribution*, $B \sim U\left(0, \frac{2}{\mu}\right)$, with $E[B] = \frac{1}{\mu}$, $E[B^2] = \frac{4}{3\mu^2}$, $V[B] = \frac{1}{3\mu^2}$, LST $\widetilde{B}(s) = \frac{\mu}{2s}\left(1 - e^{-\frac{2s}{\mu}}\right)$, and $CV^2 = \frac{1}{3}$.

Note again that all these distributions have an equal mean $E[B] = \frac{1}{\mu}$ but a different variance, enabling us to investigate the effect of the randomness of the service time on the system's performance.

6.1. Mean Number of Customers, E[L]

Let $E[L_{dis}]$ denote the mean number of customers in the system with service time distribution dis. We set $\lambda = 8$, $\mu = 10$, $\alpha = 18$, and $\beta = 20$. Figure S1 (in the Supplementary Materials File) depicts $E[L_{dis}]$ as a function of $\gamma = \{1, 2, \ldots, 200\}$ for $dis \in \{Gamma(\gamma),\ Deter,\ Unif\}$. The solid blue, dashed red, and dash-dotted green curves refer to gamma, deterministic, and uniform distributions, respectively. It is observed that $E[L_{dis}]$ decreases as a function of γ, demonstrating that the mean number of customers decreases when the variance decreases (γ increases). As expected, when $\gamma \to \infty$, $E[L_{Gamma(\gamma)}]$ converges to $E[L_{Deter}]$. We also note that $E[L_{Gamma(\gamma=3)}] = E[L_{Unif}]$.

To further investigate the impact of the different distributions on $E[L]$, we let

$$\xi = \frac{E[L_{Exp}]}{E[L_{Deter}]},\ \omega = \frac{E[L_{Exp}]}{E[L_{Unif}]},\ \psi = \frac{E[L_{Unif}]}{E[L_{Deter}]}$$

be the ratios of the mean number of customers under the different distributions. Note that $E[L_{Exp}] = E[L_{Gamma(\gamma=1)}]$.

Figures 2 and 3 depict the ratios ξ, ω, and ψ as a function of α and β by the solid blue, dashed red, and dash-dotted green curves, respectively. It follows that ξ, ω, and ψ are increasing in α (decreasing mean patience time), and decreasing in β (increasing mean orbit time). Combining these two observations, we conclude that, for a given μ, as customers spend more time in orbit, a higher improvement in $E[L]$ is achieved by decreasing the service time variance.

Similarly, Figure S2 investigates the effect of changing the parameter $\mu (\mu > \lambda = 8)$ on the ratios ξ, ω, and ψ, and illustrates that the impact of the variance on $E[L_{dis}]$ intensifies as μ decreases (i.e., a higher mean service time). Figure S2 further shows that, for a relatively short mean service time ($\mu > 40$), the ratios ξ, ω, and ψ approach 1, that is, the same $E[L_{dis}]$ for all distributions. Specifically, for high μ, the service time distribution has a negligible effect on $E[L_{dis}]$; thus, the mean number of customers is mainly affected by the patience and orbit times.

6.2. Customer's Mean Sojourn Time Excluding Orbit, E[D]

For each distribution, let $E[D_{dis}]$ be the mean sojourn time of a customer in the system, ignoring the orbit time. According to Little's law, $E[D_{dis}] = E[L_{dis}]/\lambda$. Figures 4 and 5 depict $E[D_{dis}]$ as a function of α and μ, for $dis \in \{Exp,\ Deter,\ Unif\}$ by the solid blue, dashed red, and dash-dotted curves, respectively. Similarly, Figure S3 presents $E[D_{dis}]$ as a function of β. The figures emphasize that $E[D_{dis}]$ is decreasing in α and in μ, while it is increasing in β. Thus, a lower mean patience time (a higher α) reduces $E[D_{dis}]$ (Figure 4), and a lower mean orbit time (a higher β) increases $E[D_{dis}]$ (Figure S3). The latter result is explained as follows: each customer, during her/his service processing, practically partitions the latter time into two parts: time inside the system and time in orbit. As much as β is higher (i.e., each orbit time is shorter), the customer spends more time inside the system during her/his service processing. As expected, $E[D_{dis}]$ is decreasing in μ, approaching zero when $\mu \to \infty$ (see Figure 5). Figures 4, 5 and S3 further demonstrate that $E[D_{Exp}] > E[D_{Unif}] > E[D_{Det}]$, emphasizing the increasing impact of the service time variance on the *net* sojourn time while the service continues. Evidently, the higher the variability, the longer is the mean sojourn time. Note that a higher service time variability also yields a higher mean number of customers; see Figure S1.

6.3. Mean Total Sojourn Time, E[V]

For each distribution, let $E[V_{dis}]$ denote the mean total sojourn time of a customer in the system (including orbit time). Figures 6, 7 and S4 depict $E[V_{dis}]$ as a function of α,β, and μ, respectively. The solid blue, dashed red, and dash-dotted curves refer to the exponential, deterministic, and uniform distributions, respectively. Figure 6 reveals an interesting result: $E[V_{dis}]$ is increasing in α. Thus, if the customers' mean patience time decreases (i.e., α increases such that customers go to orbit more often), they eventually spend more time in the system (in service and in orbit).

We further observe that $E[V_{dis}]$ is decreasing both in β and in μ. Interestingly, all the figures show that $E[V_{Det}] > E[V_{Unif}] > E[V_{Exp}]$, in contrast to the behavior of $E[D_{dis}]$ (see Figures 4, 5 and S3). This follows, since a smaller patience time increases the probability that the service will be completed while the customer remains in orbit.

6.4. Mean Remaining Orbit Time After Service Completion, E[H]

Let $E[H_{dis}]$ denote the mean orbit time after service completion. Figures 8, 9 and S5 depict $E[H_{dis}]$ for $dis \in \{Exp,\ Deter,\ Unif\}$ by the solid blue, dashed red, and dash-dotted curves, respectively. It is observed that $E[H_{dis}]$ is increasing in α; as the patience time decreases, the remaining orbit time increases. As expected, $E[H_{dis}]$ is decreasing in β and in μ. Furthermore, consistent with the conclusions regarding $E[V_{dis}]$ and in contrast to the results regarding $E[D_{dis}]$, we have $E[H_{Det}] > E[H_{Unif}] > E[H_{Exp}]$.

6.5. Mean Number of Orbits, E[O]

Let $E[O_{dis}]$ denote the mean number of the customer's orbits until departure. Figures 10–12 depict $E[O_{dis}]$ for $dis \in \{Exp,\ Deter,\ Unif\}$ by the solid blue, dashed red, and dash-dotted curves, respectively. Figure 10 shows that $E[O_{dis}]$ increases with α (shorter mean patience time), but the increase is bounded from above. Figure 11 demonstrates that $E[O_{dis}]$ increases with β (shorter mean orbit time). This result is explained as follows: a shorter time for each orbit excursion increases the probability of going to orbit again. Figure 12 illustrates that $E[O_{dis}]$ decreases with μ, since a shorter service time leads to fewer opportunities to go to orbit during the service time. It is also observed that, for low values of β (in particular $\beta \leq 10$), $E[O_{Unif}] > E[O_{Exp}]$. However, for higher values of β, the order changes and $E[O_{Exp}] > E[O_{Unif}]$ (see Figure 11). Additionally, according to Figure 12, for low values of μ (particularly, $\mu \leq 15$), $E[O_{Exp}] > E[O_{Unif}]$, but for higher values of μ (shorter mean service time), the opposite is true: $E[O_{Unif}] > E[O_{Exp}]$. Thus, the service time variance affects the mean number of orbits in opposite directions. Yet, when the variance becomes very small, its effect diminishes and all $E[O_{dis}]$ values become equal.

Note that the probability that a customer goes to orbit at least once equals $P(T < B)$. For example, when $B \sim Exp(10)$ and $\alpha = 18$, $P(T < B) = \frac{\alpha}{\alpha+\mu} = 0.64$.

6.6. Main Conclusions

- The mean number of customers in the system decreases when the variance of the service time decreases.
- For gamma distributed service times, the mean number of customers decreases with γ, converging to that for the deterministic distribution.
- As customers spend more time in orbit, decreasing the service time variance yields a higher reduction in the mean number of customers in the system.

- As α increases, that is, as customers are less patient and go more often to orbit, the mean total sojourn time increases where the orbit time occupies a higher proportion of the total time.
- Conversely, as β increases, that is, as the mean orbit time decreases, the mean total sojourn time decreases. Here, the service time constitutes a higher component.
- Evidently, the faster the rate of service, the shorter the sojourn times (with or without orbit).
- Ignoring the orbit time, a higher service variability (e.g., an exponential service) increases the sojourn time in the system. However, when considering the total sojourn time, it is interesting to observe that the added randomness of the orbit time balances the randomness of the service time. Consequently, a system with a higher service variance (e.g., an exponential service) is more efficient.
- Shorter patient times, shorter orbit times, and longer service times increase the mean number of orbits until departure.
- The service time variance may affect the mean number of orbits in opposite directions. Yet, when the variance becomes very small, its effect diminishes and all $E[O_{dis}]$ values become equal.

7. Summary and Further Research

This paper analyzes an M/G/1 system in the context of customers' repeated "orbit while in service" policies. Under this frequently encountered policy, each customer whose service has begun waits for a random "patience time" T and then, if the service is still in progress, goes to orbit for a random "orbit time" X. If, on return from orbit, the customer's service remains in progress, s/he draws another patience time T and waits in the system to determine whether the service has been completed within that time. If so, the customer leaves the system immediately. Otherwise, another orbit time X is drawn, and so on, until the service is completed—either while the customer is waiting in the system or when s/he is in orbit. The server continues rendering service for this customer even while s/he is in orbit. By encouraging customers to implement this policy, their "waiting for service" time may be used more efficiently.

Various system performance measures are derived using the SVT and LSTs. We first obtain the PGFs of L, the number of customers in the system (not including those in orbit), which allows us to calculate $E[L]$. We then consider V, the total sojourn time of a customer in the system, starting from the moment s/he starts service until departure, and derive the corresponding LST. We further define H as the time a customer remains in orbit after her/his service has been completed and derive its LST. We then consider $E[D] = E[L]/\lambda$, the customer's mean sojourn time in the system, not including orbit. Finally, we obtain $E[O]$, a customer's mean number of orbits before departure. We consider a range of service time distributions and conduct a comparison analysis for the above performance measures. Specifically, we consider the family of gamma probability distribution functions, spanning the range between the exponential and deterministic distributions, as well as the uniform distribution. This investigation reveals that a customer's mean total sojourn time (inside the system and in orbit) achieves the highest values when the service time is deterministic and the lowest values when the service time is exponentially distributed. However, the lowest value of the customer's mean sojourn time inside the system (excluding orbit) is achieved when the service time is deterministic and the highest when the service time is exponentially distributed.

For future research, we propose considering and analyzing "orbit while in service" scenarios with general patience time and general orbit times rather than exponentially distributed times. Another direction for future investigation is to consider varying distributions for orbit times with decreasing means. Similarly, one could consider successive patience times with decreasing or increasing means (in reality, some customers become increasingly less patient while others are willing to remain longer in the system, assuming that the service time distribution is of an increasing failure rate). Finally, it could be interesting to study the "orbit while in service" model with priorities.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math12233722/s1, Figure S1: $E[L_{dis}]$ for gamma (γ = {1,2,...,200}), deterministic, and uniform distributions. Figure S2: ξ, ω, and ψ as a function of μ. Figure S3: $E[D_{dis}]$ as a function of β. Figure S4: $E[V_{dis}]$ as a function of μ. Figure S5: $E[H_{dis}]$ as a function of β.

Author Contributions: Conceptualization, G.H., Y.B. and U.Y.; methodology, G.H., Y.B. and U.Y.; software, G.H.; validation, G.H., Y.B. and U.Y.; formal analysis, G.H., Y.B. and U.Y.; investigation, G.H., Y.B. and U.Y.; writing—original draft preparation, G.H., Y.B. and U.Y.; writing—review and editing, G.H., Y.B. and U.Y.; visualization, G.H., Y.B. and U.Y. All authors have read and agreed to the published version of the manuscript.

Funding: Uri Yechiali's work was supported by the Israel Science Foundation, grant number 1968/23.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Materials. Further inquiries can be directed to the corresponding author.

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

References

1. Templeton, J.G.C.; Falin, G.I. *Retrial Queues*; Taylor & Francis: Abingdon, UK, 2023; ISBN 1351418661.
2. Artalejo, J.R.; Gómez-Corral, A. *Retrial Queueing Systems: A Computational Approach*; Springer: Berlin/Heidelberg, Germany, 2008; ISBN 3540787240.
3. Gómez-Corral, A.; Phung-Duc, T. Retrial Queues and Related Models. *Ann. Oper. Res.* **2016**, *247*, 1–2. [CrossRef]
4. Hanukov, G.; Anily, S.; Yechiali, U. Ticket Queues with Regular and Strategic Customers. *Queueing Syst.* **2020**, *95*, 145–171. [CrossRef]
5. Hanukov, G.; Yechiali, U. Orbit While in Service. *Oper. Res.* **2024**, *24*, 18. [CrossRef]
6. Hanukov, G. A Queueing-Inventory System with a Repeated-Orbit Policy during the Service. *Ann. Oper. Res.* **2023**, 1–33. [CrossRef]
7. Kosten, L. Stochastic Theory of Service Systems. *Oper. Res. Q.* **1974**, *25*, 645. Available online: https://cir.nii.ac.jp/crid/1130000796048542720 (accessed on 1 October 2024).
8. Knessl, C.; Matkowsky, B.J.; Schuss, Z.; Tier, C. An Integral Equation Approach to the M/G/2 Queue. *Oper. Res.* **1990**, *38*, 506–518. [CrossRef]
9. Henderson, W. Alternative Approaches to the Analysis of the M/G/1 and G/M/1 Queues. *J. Oper. Res. Soc. Jpn.* **1972**, *15*, 92–101.
10. Klimko, E.M.; Neuts, M.F. The Single Server Queue in Discrete Time-numerical Analysis II. *Nav. Res. Logist. Q.* **1973**, *20*, 305–319. [CrossRef]
11. DO, L.E. The Discrete-Time Single-Server Queue with Time-Inhomogeneous Compound Poisson Input and General Service Time Distribution. *J. Appl. Prob.* **1978**, *15*, 601.
12. Barron, Y. A Replenishment Inventory Model with a Stock-Dependent Demand and Age–Stock-Dependent Cost Functions in a Random Environment. *Asia-Pac. J. Oper. Res.* **2022**, *39*, 2150035. [CrossRef]
13. Cohen, J.W. *The Single Server Queue*; Elsevier: Amsterdam, The Netherlands, 1982.
14. Kerner, Y. The Conditional Distribution of the Residual Service Time in the M n/G/1 Queue. *Stoch. Model.* **2008**, *24*, 364–375. [CrossRef]
15. Kulkarni, V.G.; Choi, B.D. Retrial Queues with Server Subject to Breakdowns and Repairs. *Queueing Syst.* **1990**, *7*, 191–208. [CrossRef]
16. Kerner, Y.; Shmuel-Bittner, O. Strategic Behavior and Optimization in a Hybrid M/M/1 Queue with Retrials. *Queueing Syst.* **2020**, *96*, 285–302. [CrossRef]
17. Fiems, D. Retrial Queues with Constant Retrial Times. *Queueing Syst.* **2023**, *103*, 347–365. [CrossRef]
18. Dimitriou, I. A Single Server Retrial Queue with Event-Dependent Arrival Rates. *Ann. Oper. Res.* **2023**, *331*, 1053–1088. [CrossRef]
19. Jeganathan, K.; Harikrishnan, T.; Lakshmanan, K.; Melikov, A.; Sztrik, J. Modeling of Junior Servers Approaching a Senior Server in the Retrial Queuing-Inventory System. *Mathematics* **2023**, *11*, 4581. [CrossRef]
20. Hu, K.; Allon, G.; Bassamboo, A. Understanding Customer Retrials in Call Centers: Preferences for Service Quality and Service Speed. *Manuf. Serv. Oper. Manag.* **2022**, *24*, 1002–1020. [CrossRef]
21. Zhang, Y.; Wang, J. Managing Retrial Queueing Systems with Boundedly Rational Customers. *J. Oper. Res. Soc.* **2023**, *74*, 748–761. [CrossRef]
22. Xu, J.; Liu, L.; Wu, K. Analysis of a Retrial Queueing System with Priority Service and Modified Multiple Vacations. *Commun. Stat.-Theory Methods* **2022**, *52*, 6207–6231. [CrossRef]

23. Lee, S.W.; Kim, B.; Kim, J. Analysis of the Waiting Time Distribution in M/G/1 Retrial Queues with Two Way Communication. *Ann. Oper. Res.* **2022**, *310*, 505–518. [CrossRef]
24. Avrachenkov, K.; Yechiali, U. Retrial Networks with Finite Buffers and Their Application to Internet Data Traffic. *Probab. Eng. Inf. Sci.* **2008**, *22*, 519–536. [CrossRef]
25. Yechiali, U. Queues with System Disasters and Impatient Customers When System Is Down. *Queueing Syst.* **2007**, *56*, 195–202. [CrossRef]
26. Liu, H.-L.; Li, Q.-L. Matched Queues with Flexible and Impatient Customers. *Methodol. Comput. Appl. Probab.* **2023**, *25*, 4. [CrossRef]
27. Aalto, S. Whittle Index Approach to Multiserver Scheduling with Impatient Customers and DHR Service Times. *Queueing Syst.* **2024**, *107*, 1–30. [CrossRef]
28. Manitz, M.; Piehl, M.-P. A Fast Staffing Algorithm for Multistage Call Centers with Impatient Customers and Time-Dependent Overflow. *Cent. Eur. J. Oper. Res.* **2024**, *32*, 763–791. [CrossRef]
29. Lv, S.; Li, F.; Li, J. The M/M/c Retrial Queueing System with Impatient Customers and Server Working Breakdown. *IAENG Int. J. Appl. Math.* **2024**, *54*, 1499.
30. Altman, E.; Yechiali, U. Analysis of Customers' Impatience in Queues with Server Vacations. *Queueing Syst.* **2006**, *52*, 261–279. [CrossRef]
31. Altman, E.; Yechiali, U. Infinite-Server Queues with System's Additional Tasks and Impatient Customers. *Probab. Eng. Inf. Sci.* **2008**, *22*, 477–493. [CrossRef]
32. Inoue, Y.; Ravner, L.; Mandjes, M. Estimating Customer Impatience in a Service System with Unobserved Balking. *Stoch. Syst.* **2023**, *13*, 181–210. [CrossRef]
33. Bassamboo, A.; Randhawa, R.; Wu, C. Optimally Scheduling Heterogeneous Impatient Customers. *Manuf. Serv. Oper. Manag.* **2023**, *25*, 1066–1080. [CrossRef]
34. Sharma, S.; Kumar, R.; Soodan, B.S.; Singh, P. Queuing Models with Customers' Impatience: A Survey. *Int. J. Math. Oper. Res.* **2023**, *26*, 523–547. [CrossRef]
35. Levy, Y.; Yechiali, U. Utilization of Idle Time in an M/G/1 Queueing System. *Manag. Sci.* **1975**, *22*, 202–211. [CrossRef]
36. Levy, Y.; Yechiali, U. An M/M/s Queue with Servers' Vacations. *INFOR Inf. Syst. Oper. Res.* **1976**, *14*, 153–163. [CrossRef]
37. Takagi, H. Queueing Analysis: A Foundation of Performance Evaluation. In *Vacation and Priority Systems*; Elsevier: Amsterdam, The Netherlands, 1991; Volume 1.
38. Tian, N.; Zhang, Z.G. *Vacation Queueing Models: Theory and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006; Volume 93, ISBN 0387337237.
39. Barron, Y. A Threshold Policy in a Markov-Modulated Production System with Server Vacation: The Case of Continuous and Batch Supplies. *Adv. Appl. Probab.* **2018**, *50*, 1246–1274. [CrossRef]
40. Niranjan, S.P.; Devi Latha, S.; Mahdal, M.; Karthik, K. Multiple Control Policy in Unreliable Two-Phase Bulk Queueing System with Active Bernoulli Feedback and Vacation. *Mathematics* **2023**, *12*, 75. [CrossRef]
41. Liu, T.-H.; Chiou, K.-C.; Chen, C.-M.; Chang, F.-M. Multiserver Retrial Queue with Two-Way Communication and Synchronous Working Vacation. *Mathematics* **2024**, *12*, 1163. [CrossRef]
42. Khan, I.E.; Paramasivam, R. Reduction in Waiting Time in an M/M/1/N Encouraged Arrival Queue with Feedback, Balking and Maintaining of Reneged Customers. *Symmetry* **2022**, *14*, 1743. [CrossRef]
43. Melikov, A.; Aliyeva, S.; Sztrik, J. Analysis of Instantaneous Feedback Queue with Heterogeneous Servers. *Mathematics* **2020**, *8*, 2186. [CrossRef]
44. Melikov, A.; Chakravarthy, S.R.; Aliyeva, S. A Retrial Queueing Model with Feedback. *Queueing Models Serv. Manag.* **2023**, *6*, 63–95.
45. Mytalas, G.C.; Zazanis, M.A. Performance Analysis for Bernoulli Feedback Queues Subject to Disasters: A System with Batch Poisson Arrivals under a Multiple Vacation Policy. *Qual. Technol. Quant. Manag.* **2023**, *20*, 113–146. [CrossRef]
46. Ross, S.M. *Stochastic Processes*; Wiley: New York, NY, USA, 1996; Volume 2.