Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An M/G/1 Queueing Model with Bernoulli Catastrophes and Repair Times

Rein Nobel

r.d.nobel@vu.nl

Vrije Universiteit Amsterdam

Tjalling Veenstra

Vrije Universiteit Amsterdam

Research Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11351424/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An $M/G/1$ Queueing Model with Bernoulli Catastrophes and Repair Times

Rein Nobel & Tjalling Veenstra
r.d.nobel@vu.nl ; s.t.veenstra@vu.nl

Department of Operations Analytics,
Vrije Universiteit Amsterdam.

Abstract

We study a single-server queueing model of the $M/G/1$ type with so-called Bernoulli catastrophes. Customers arrive according to a Poisson process with rate λ and are queued in case they find the server busy or under repair. The queue is served in FIFO order. Each customer requires a *Coxian*-2 distributed service time with parameters (b, μ_1, μ_2). To this $M/C_2/1$ queueing model a Poisson stream of catastrophes is added with rate γ. Upon occurrence of a catastrophe with probability α the queueing system (the queued customers plus the customer in service) is emptied completely, and with probability $1-\alpha$ a catastrophe only eliminates the customer in service (a so-called negative arrival). We consider two versions: (i) without repair times, i.e. upon occurrence of a catastrophe the server can immediately start a new service when a customer is or will be present, and (ii) with repair times, i.e. upon occurrence of a catastrophe the server starts an exponential repair time with rate ν. For both versions we study the steady-state behaviour by deriving the (partial) probability generating functions (PGFs) of the limiting probabilities of the system-states. From these PGFs explicit expressions for the probabilities of the system states will be deduced, as well as several other characteristics of the system, such as the long-run fraction of time that the server is idle, busy or under repair, the steady-state queue-length distribution, the waiting-time distribution of a so-called 'successful customer' (i.e. a customer who will reach the server to start his service), the long-run fraction of these successful customers, and many more. Numerical results will give us further insight into the behaviour of this $M/G/1$ queueing model with Bernoulli catastrophes and repair times.

1 Introduction

Over the past five decades classical $GI/G/c$ type queueing models have been extended in many ways to incorporate a wide range of service interruptions or service rate changes, often accompanied by the departure or elimination of customers from the system. Service interruptions can be initiated by the server or by events coming from the outside world. Models with the option for so-called server *vacations* are of the first category: the server decides to interrupt the service or to continue service at a lower rate, for instance at the end of a busy cycle, and goes on a (working) vacation during which arriving customers usually have to wait (longer) until the server is available again or returns to the original service rate. Many variants of these so-called *vacation queueing models* have been studied since the seventies of the last century, see e.g. Nashuo Tian et al. [19] for a comprehensive monograph. Vacation queueing models can be extended further by introducing *impatience* of customers waiting in the queue (or served at a lower rate) during a vacation. We refer to Eitan Altman et al. [1], Veena Goswami [8], and

Ahakir Majid [16], Vijayashree et al. [21], and the references therein for examples of (working) vacation models with impatient customers.

The second category of queueing models with service interruptions is characterised by events beyond control of the server, e.g. a breakdown of a machine (server) in a production facility, when the service interruption often also leads to the elimination (destruction) of one or more customers from the system. Such events can take several forms. When only one customer, e.g. the customer in service, is eliminated the damaging events are usually modelled by an arrival stream of so-called *negative customers*: upon arrival of a negative customer one customer is eliminated from the system, see e.g. Gelenbe et al. [7] and Harrison et al. [9] and the references therein. Of course, it is possible that upon the occurrence of such an outside event all customers are eliminated from the system, e.g. the occurrence of a virus in a computer or communication system which destroys not only the job(s) in service but also all the jobs waiting to be processed. In that case we speak of a *disaster* or *catastrophe* (also terms like 'mass exodus', 'queue flushing' and 'stochastic clearing' are used). Also these queueing models with catastrophes have been widely studied for many decades. We mention Gautam Jain et al. [10], B. Krishna Kumar et al. [12], A. di Crescenzo et al. [3], Mustafa Demircioglu et al. [4] (a discrete-time model with catastrophes), Bo Keun Kim et al. [11], and Young Liu et al. [15], and the reference therein. Often these models are extended with a *repair time* as a requirement before the server can restart his normal services, see e.g. Krishna Kumar et al. [13], Sundar Rajan Balasubramanian et al. [2], Bo Keun Kim et al. [11], Krishna Kumar et al. [14], Ramachandran Remya et al. [17], and Thilagavathy et al. [18], and the references in these papers. Not surprisingly, also queueing models with catastrophes and repair times have been extended further to incorporate impatience of customers waiting in the queue while the server is under repair. We mention Uri Yechiali [22], Spiros Dimou et al. [5], and Antonis Economou et al. [6] as a few examples.

In Krishna Kumar et al. [13] the authors study an $M/M/1$ queue with catastrophes and repair times, but in their model all customers arrived while the server is under repair are lost. In other papers, e.g. Yechiali [22], catastrophes are disabled while the server is under repair. Further, in most papers mentioned above the service times follow an exponential distribution, which impedes any observation of differences in the performance measures for regular versus very irregular service times to name just two extremes, and the authors usually choose to analyse either a model with negative customers or a model with catastrophes. Seldom we see a paper with both types of service interruptions. In Thilagavathy et al. [18] the authors study a $MAP/PH/1$ queueing system with negative arrivals, catastrophic failures, and repairs. They calculate algorithmically the steady-state probability vector of the Quasi-Birth-Death matrix of this model, formulate several performance measures in terms of these (approximated) steady-state probabilities (of course, not in the form of explicit expressions), but the analysis (and also the model itself) is not very clear, and their results do not give much insight into the sensitivities of the performance measures for varying parameter values (numerical results are presented only for the mean number of customers in the system for different renege rates).

Against the background of the above (albeit limited) literature overview, we will present an $M/G/1$ queueing model with Coxian-2 distributed service times (so we can vary the squared coefficient of variation, *ceteris paribus*), an external Poisson arrival stream of so-called *Bernoulli catastrophes*, i.e. catastrophes come in two forms: either as disasters (from now on, *complete catastrophes*) or as negative arrivals (from now on, *local catastrophes*), and exponentially distributed repair times. While the server is under repair customers continue to arrive and queue at the system (in this paper we do not incorporate impatient customers), and the Poisson stream of (complete) catastrophes is not disabled. For this model we will derive explicit expressions for the steady-state probabilities of the queue length distribution, the mean queue length, the mean work load (virtual waiting time), the overall mean waiting-time, and the mean waiting time of a so-called successful customer, i.e. a customer who will not be eliminated before

his arrival at the server.

In Section 2 we will present the steady-state analysis of a (simplified) version of this model without repair times, i.e. after a catastrophe (local or complete) the server will restart his services as soon as a customer is present in the system. We will derive the (partial) probability generating functions (PGFs) of the system-states. It turns out that the denominator of these PGFs is a cubic polynomial, and hence we can find explicit expressions for the zeroes of the denominator using Girolamo Cardano's formulae, which enable us to give explicit expressions for the steady-state probabilities of the queue length. Further, the PGFs will be used to find the Laplace-Stieltjes Transform (LST) of the waiting-time distribution of a successful customer, and also the fraction of customers who leave the system after service completion will be calculated.

In Section 3 we will present the steady-state analysis of the model with repair times. Again the PGFs of the system-states are derived, and now the denominator is a quartic polynomial, hence we can still give explicit expressions for the zeroes of this denominator, using the formulae of Ludovico Ferrari for the solution of a quartic equation. Using these zeroes we present explicit expressions for the steady-state probabilities of the queue length. The rest of Section 3 follows the same lines as Section 2: we calculate the same performance measures, but all calculations are much more complicated. Finally, in Section 4 we present many numerical results for the simplified model without repair times, as well as for the model with repair times. An at first sight counter-intuitive result of our numerical investigations is that in both models we do not observe the monotonicity properties with respect to the squared coefficient of variation of the service time, which are so characteristic for standard $M/G/1$ queues (the Pollaczek-Khinchine formula!): in our models with Bernoulli catastrophes the important performance measures (the mean queue length, the mean work load, the overall mean waiting time, the mean waiting time of a successful customer, to name a few) are not monotone functions of the squared coefficient of variation of the service time! We reckon these non-monotonicity findings among the most striking insights obtained from the steady-state analysis of these models, and in Section 4 we present a thorough explanation for this contrast with the 'Pollaczek-Khinchine monotonicity' of the standard $M/G/1$ queue.

In a subsequent paper we hope to incorporate impatience for the customers who remain in the queue while the server is under repair.

2 The steady-state analysis of the simplified version

In this section we will analyse the $M/C_2/1$ queueing model with Bernoulli catastrophes without repair times. Customers arrive at a single server according to a Poisson process with rate λ. Every customer requires a Coxian-2 distributed service time with parameters b, μ_1 and μ_2, i.e. every service starts with a first exponential phase with rate μ_1, and with probability b this phase is followed by a second exponential phase with rate μ_2. Customers who upon arrival find the server busy (or under repair; see Section 3) are placed in a queue which is served in FIFO order. This $M/C_2/1$ queueing model is extended with a Poisson stream of catastrophes with rate γ as long as the server is busy (or under repair). Upon occurrence of a catastrophe with probability α all customers are removed (i.e. the customers waiting in the queue plus the customer in service), termed as a *complete catastrophe*, and with probability $1-\alpha$ only the customer in service is removed, a so-called negative arrival, here termed as a *local catastrophe*. Removed customers do not play any future role in the system.

To study the steady-state behaviour of this model we introduce the continuous-time Markov Chain (CTMC) $\{(C(t), Q(t)), t \geq 0\}$, where $C(t)$ describes the state of the server at time t, and $Q(t)$ the

number of customers waiting in the queue. More precisely, we define

$$C(t) = \begin{cases} 0 & \text{when the server is idle at time } t, \\ 1 & \text{when the server is busy with a first service phase at time } t, \\ 2 & \text{when the server is busy with a second service phase at time } t, \\ 3 & \text{when the server is under repair at time } t, \end{cases}$$

and $Q(t)$ the number of queued customers at time t. Notice that in this section $C(t) = 3$ does not occur because for the moment we consider the simplified version of the model without repair times.

Due to the Poisson stream of catastrophes stability is not an issue, and we can define the limiting probabilities without any restrictions on the value of the parameters of the model. So, we can introduce

$$\begin{aligned} p_{00} &:= \lim_{t\to\infty} \mathbb{P}(C(t) = 0; Q(t) = 0) \\ p_{in} &:= \lim_{t\to\infty} \mathbb{P}(C(t) = i; Q(t) = n), \quad i = 1, 2, 3;\ n = 0, 1, 2, \ldots, \end{aligned}$$

and the (partial) probability generating functions (PGFs)

$$P_i(z) := \sum_{n=0}^{\infty} p_{in} z^n, \quad i = 1, 2, 3.$$

To find closed form expressions for these PGFs we write down the balance equations for the simplified model,

$$\lambda p_{00} = (1-b)\mu_1 p_{10} + \mu_2 p_{20} + \alpha\gamma(1 - p_{00}) + (1-\alpha)\gamma(p_{10} + p_{20}) \tag{1}$$

$$(\lambda + \gamma + \mu_1)p_{10} = \lambda p_{00} + (1-b)\mu_1 p_{11} + \mu_2 p_{21} + (1-\alpha)\gamma(p_{11} + p_{21}) \tag{2}$$

$$(\lambda + \gamma + \mu_1)p_{1n} = \lambda p_{1,n-1} + (1-b)\mu_1 p_{1,n+1} + \mu_2 p_{2,n+1} + (1-\alpha)\gamma(p_{1,n+1} + p_{2,n+1}), \quad n = 1, 2, \ldots \tag{3}$$

$$(\lambda + \gamma + \mu_2)p_{2n} = \lambda p_{2,n-1} + b\mu_1 p_{1,n}, \quad n = 0, 1, 2, \ldots. \tag{4}$$

Using the standard procedures (we skip the details), from Eqs. (1)–(4) we find

$$P_1(z) = \frac{(\lambda(1-z) + \gamma + \mu_2)(\alpha\gamma - (\lambda(1-z) + \alpha\gamma)p_{00})}{[z(\lambda(1-z) + \gamma + \mu_1) - ((1-b)\mu_1 + (1-\alpha)\gamma)](\lambda(1-z) + \gamma + \mu_2) - b\mu_1(\mu_2 + (1-\alpha)\gamma)}, \tag{5}$$

$$P_2(z) = \frac{b\mu_1}{\lambda(1-z) + \gamma + \mu_2} P_1(z). \tag{6}$$

To find the unknown probability p_{00} we consider the denominator $\mathcal{D}(z)$ of the expression for $P_1(z)$, i.e.

$$\mathcal{D}(z) = [z(\lambda(1-z) + \gamma + \mu_1) - ((1-b)\mu_1 + (1-\alpha)\gamma)](\lambda(1-z) + \gamma + \mu_2) - b\mu_1(\mu_2 + (1-\alpha)\gamma).$$

We see that $\mathcal{D}(z)$ is a *cubic* polynomial in z with $\mathcal{D}(0) < 0$ and $\mathcal{D}(1) > 0$, so there is a real number $z_0 \in (0, 1)$ for which $\mathcal{D}(z_0) = 0$. To show that this z_0 is the unique zero in the unit disk $|z| < 1$ we apply *Rouché's Theorem*. Let $f(z) = [z(\lambda(1-z) + \gamma + \mu_1) - ((1-b)\mu_1 + (1-\alpha)\gamma)](\lambda(1-z) + \gamma + \mu_2)$ and $g(z) = -b\mu_1(\mu_2 + (1-\alpha)\gamma)$, so $f(z) + g(z) = \mathcal{D}(z)$. Then, it can be easily checked that $f(z)$ has exactly one zero within the unit disk $|z| < 1$, because the first (quadratic) factor of $f(z)$ has two real zeroes, one in the open interval $(0, 1)$, and the other larger than 1, and the second factor has the zero $z = (\lambda + \gamma + \mu_2)/\lambda > 1$. Further, for all $z \in \Gamma = \{z \in \mathbb{C} : |z| = 1\}$ we have

$$|g(z)| = b\mu_1(\mu_2 + (1-\alpha)\gamma).$$

Next, we can find a lower bound for the quadratic factor $|z(\lambda(1-z)+\gamma+\mu_1)-((1-b)\mu_1+(1-\alpha)\gamma)|$ of $|f(z)|$ for $z \in \Gamma$, using the roots $r_1 \in (0,1)$ and $r_2 > 1$, the formulas for the sum $r_1 + r_2$ and the product $r_1 r_2$, and the triangle inequality $|x+y| \geq ||x|-|y|| \geq |x|-|y|$, as follows

$$|z(\lambda(1-z)+\gamma+\mu_1)-((1-b)\mu_1+(1-\alpha)\gamma)| = |\lambda z^2-(\lambda+\gamma+\mu_1)z+(1-b)\mu_1+(1-\alpha)\gamma| = \lambda|z-r_1||z-r_2| \geq$$

$$\lambda||z|-r_1||z-r_2| = \lambda|(1-r_1)z-(1-r_1)r_2| \geq \lambda|(1-r_1)|z|-(1-r_1)r_2| = \lambda|1-(r_1+r_2)+r_1 r_2| =$$

$$\lambda\left|1-\frac{\lambda+\gamma+\mu_1}{\lambda}+\frac{(1-b)\mu_1+(1-\alpha)\gamma}{\lambda}\right| = b\mu_1+\alpha\gamma.$$

Then we get for all $z \in \Gamma$

$$\begin{aligned} |f(z)| &\geq (b\mu_1+\alpha\gamma)|\lambda(1-z)+\gamma+\mu_2| \geq (b\mu_1+\alpha\gamma)|\lambda+\gamma+\mu_2-\lambda|z|| = \\ &\quad (b\mu_1+\alpha\gamma)(\gamma+\mu_2) > b\mu_1(\mu_2+(1-\alpha)\gamma) = |g(z)|. \end{aligned}$$

So, from Rouché's Theorem we can conclude that $\mathcal{D}(z) = f(z) + g(z)$ has a unique zero $z = z_0$ within the unit disk $|z| < 1$. As a consequence, also the numerator must vanish for $z = z_0$, and we find

$$p_{00} = \frac{\alpha\gamma}{\lambda(1-z_0)+\alpha\gamma}. \tag{7}$$

2.1 The Queue Length Distribution

With *Girolamo Cardano's* Formula we can find the zero z_0 as well as the other two zeroes, say z_1 and z_2. Then we can factorise $\mathcal{D}(z)$ as follows

$$\mathcal{D}(z) = \lambda^2(z-z_0)(z-z_1)(z-z_2),$$

and cancelling the factor $z - z_0$ in the numerator and denominator of $P_1(z)$ and $P_2(z)$ we get

$$P_1(z) = \frac{(\lambda(1-z)+\gamma+\mu_2)\left(\frac{\lambda\alpha\gamma}{\lambda(1-z_0)+\alpha\gamma}\right)}{\lambda^2(z-z_1)(z-z_2)} = \frac{\alpha\gamma(\lambda(1-z)+\gamma+\mu_2)}{\lambda(\lambda(1-z_0)+\alpha\gamma)(z-z_1)(z-z_2)} \tag{8}$$

$$P_2(z) = \frac{b\mu_1}{\lambda(1-z)+\gamma+\mu_2}P_1(z) = \frac{\alpha\gamma b\mu_1}{\lambda(\lambda(1-z_0)+\alpha\gamma)(z-z_1)(z-z_2)}. \tag{9}$$

Then, using Eq. (9) we can rewrite $P_2(z)$ as

$$P_2(z) = \frac{\alpha\gamma b\mu_1}{\lambda(\lambda(1-z_0)+\alpha\gamma)(z_2-z_1)}\left[\frac{1}{z_1-z}-\frac{1}{z_2-z}\right] = \tag{10}$$

$$\frac{\alpha\gamma b\mu_1}{\lambda(\lambda(1-z_0)+\alpha\gamma)(z_2-z_1)}\left[\frac{1}{z_1}\sum_{n=0}^{\infty}\left(\frac{z}{z_1}\right)^n-\frac{1}{z_2}\sum_{n=0}^{\infty}\left(\frac{z}{z_2}\right)^n\right]. \tag{11}$$

So we find for $n = 0, 1, 2, \ldots$

$$p_{2n} = \frac{\alpha\gamma b\mu_1}{\lambda(\lambda(1-z_0)+\alpha\gamma)(z_2-z_1)}\left[\left(\frac{1}{z_1}\right)^{n+1}-\left(\frac{1}{z_2}\right)^{n+1}\right].$$

Further we know that

$$P_1(z) = \sum_{n=0}^{\infty}p_{1n}z^n = \frac{\lambda(1-z)+\gamma+\mu_2}{b\mu_1}\cdot P_2(z) = \frac{\lambda+\gamma+\mu_2}{b\mu_1}P_2(z)-\frac{\lambda z}{b\mu_1}P_2(z), \tag{12}$$

from which we get

$$p_{1n} = \frac{\lambda+\gamma+\mu_2}{b\mu_1}p_{2n} - \frac{\lambda}{b\mu_1}p_{2,n-1}, \quad n = 0, 1, 2, \ldots$$

where we define $p_{2,-1} = 0$.

From the above results we can easily deduce the long-run average queue length L_q, because

$$L_q = \sum_{n=1}^{\infty} n(p_{1n} + p_{2n}) = P_1'(1) + P_2'(1).$$

We find by straightforward differentiation of Eqs. (8) and (9)

$$L_q = \frac{\lambda}{\alpha\gamma}\left(1 - \frac{(1-z_0)(\gamma+\mu_1)(\gamma+\mu_2)}{(\gamma+b\mu_1+\mu_2)[\lambda(1-z_0)+\alpha\gamma]}\right). \tag{13}$$

From now on we will use the random variables C and Q for the steady-state server state and the queue length, respectively. Then for the queue length probabilities, say $q_n := \mathbb{P}(Q = n)$ we get for $n = 0, 1, 2, \ldots$

$$\begin{aligned} q_n &= p_{00}\mathbf{I}_{\{n=0\}} + p_{1n} + p_{2n} = p_{00}\mathbf{I}_{\{n=0\}} + \frac{\lambda+\gamma+b\mu_1+\mu_2}{b\mu_1}p_{2n} - \frac{\lambda}{b\mu_1}p_{2,n-1} \\ &= \frac{\alpha\gamma}{\lambda(1-z_0)+\alpha\gamma}\left\{\mathbf{I}_{\{n=0\}} + \frac{\lambda(1-z_1)+\gamma+b\mu_1+\mu_2}{\lambda z_1(z_2-z_1)}\left(\frac{1}{z_1}\right)^n - \frac{\lambda(1-z_2)+\gamma+b\mu_1+\mu_2}{\lambda z_2(z_2-z_1)}\left(\frac{1}{z_2}\right)^n\right\}. \end{aligned}$$

To conclude this paragraph, using the normalization condition and Eq. (6), the long-run fractions of time that the server is in the different states are

$$\begin{aligned} \mathbb{P}(C=0) &= p_{00}, \\ \mathbb{P}(C=1) &= P_1(1) = \frac{\gamma+\mu_2}{\gamma+b\mu_1+\mu_2}(1-p_{00}), \\ \mathbb{P}(C=2) &= P_2(1) = \frac{b\mu_1}{\gamma+b\mu_1+\mu_2}(1-p_{00}). \end{aligned}$$

2.2 The Waiting Time distribution(s)

Next, we will discuss the steady-state waiting-time distribution, i.e. the time that a typical customer has to stay in the queue before the start of his service. Firstly, we remark that using Little's Law is not very useful here because many queued customers will never start their service due to the occurrence of a complete catastrophe while they stay in the queue. Hence, L_q/λ is the mean waiting time calculated over *all* customers in the queue, so included the customers prematurely eliminated due to a complete catastrophe. A much more interesting performance measure is the waiting time of the customers who will arrive 'safely' at the server. These so-called *successful customers* only might have been confronted during their stay in the queue with one or more local catastrophes (which do not hurt them), but not with any complete catastrophe (which would have eliminated them from the queue). To calculate the waiting-time distribution of a successful customer we introduce the random variable D for the waiting time of a typical customer (often called the *tagged* customer) as if complete catastrophes are temporarily neglected. D is the so-called *virtual waiting time*, i.e. the amount of work in the system at the time of arrival (also called the *workload*), or in other words, the waiting time as 'foreseen' by a customer upon arrival, before a complete catastrophe might eliminate him from the system. Take a typical customer who upon arrival 'sees' the system in state $(C, Q) = (1, n)$. Then his virtual waiting time expires after

a mixture of $n+1$ Coxian-2 service completions and local catastrophes. Let $S_j \sim \text{Cox}(b, \mu_1, \mu_2)$ be the service time of the j-th customer observed by the typical customer at the server (here the first customer is the customer at the server at the time of arrival, the second is the customer at the head of the queue at the time of arrival, et cetera), and introduce the generic random variables $E(\theta) \sim \text{Exp}(\theta)$ for $\theta > 0$. Then we can write

$$\mathbb{P}(D > t | C = 1; Q = n) \;=\; \mathbb{P}\left(\sum_{j=1}^{n+1} S_j^* > t\right), \tag{14}$$

where we define $S_j^* := \min\{S_j, E((1-\alpha)\gamma)\}$. So, we have to calculate the distribution of the S_j^*. Recall that for $\mu_1 \neq \mu_2$ the density of the Coxian(b, μ_1, μ_2)-distributed service times S_j is given by

$$f_S(t) = \left(1 - \frac{b\mu_1}{\mu_1 - \mu_2}\right) \mu_1 \mathrm{e}^{-\mu_1 t} + \frac{b\mu_1}{\mu_1 - \mu_2} \mu_2 \mathrm{e}^{-\mu_2 t}.$$

So, for the density of the S_j^* we easily find

$$f_{S^*}(t) = \left(1 - \frac{b\mu_1}{\mu_1 - \mu_2}\right) (\mu_1 + (1-\alpha)\gamma) \mathrm{e}^{-(\mu_1 + (1-\alpha)\gamma)t} + \frac{b\mu_1}{\mu_1 - \mu_2} (\mu_2 + (1-\alpha)\gamma) \mathrm{e}^{-(\mu_2 + (1-\alpha)\gamma)t},$$

and hence for the Laplace-Stieltjes Transform (LST)

$$\hat{\mathcal{F}}_{S^*}(s) = \left(1 - \frac{b\mu_1}{\mu_1 - \mu_2}\right) \frac{\mu_1 + (1-\alpha)\gamma}{s + \mu_1 + (1-\alpha)\gamma} + \frac{b\mu_1}{\mu_1 - \mu_2} \cdot \frac{\mu_2 + (1-\alpha)\gamma}{s + \mu_2 + (1-\alpha)\gamma}.$$

Similarly, when the typical customer 'sees' upon arrival the system-state $(C, Q) = (2, n)$ we have

$$\mathbb{P}(D > t | C = 2; Q = n) \;=\; \mathbb{P}\left(E(\mu_2 + (1-\alpha)\gamma) + \sum_{j=2}^{n+1} S_j^* > t\right). \tag{15}$$

Summarising the above results, we find for the LST of the steady-state virtual waiting time D of a typical customer, using the *Law of Total Probability* and the PASTA property

$$\hat{\mathcal{D}}(s) \;=\; p_{00} + \hat{\mathcal{F}}_{S^*}(s) P_1\left(\hat{\mathcal{F}}_{S^*}(s)\right) + \frac{\mu_2 + (1-\alpha)\gamma}{s + \mu_2 + (1-\alpha)\gamma} P_2\left(\hat{\mathcal{F}}_{S^*}(s)\right). \tag{16}$$

Notice that $\mathbb{P}(D = 0) = p_{00}$.

Once we know the distribution of the virtual waiting time D we can calculate the long-run fraction of successful customers, i.e. the fraction of customers who successfully reach the server. Because the Poisson stream of catastrophes with rate γ is randomly split into two *independent* Poisson streams of local catastrophes and complete catastrophes with rates $(1-\alpha)\gamma$ and $\alpha\gamma$, respectively, we find for a typical customer

$$\begin{aligned}\mathbb{P}(\text{customer successfully starts his service}) = \mathbb{P}(E(\alpha\gamma) > D) &= \int_0^\infty \mathbb{P}(E(\alpha\gamma) > D \mid D = t)\, f_D(t)\, \mathrm{d}t \\ &= \int_0^\infty \mathrm{e}^{-\alpha\gamma t} f_D(t)\, \mathrm{d}t = \hat{\mathcal{D}}(\alpha\gamma).\end{aligned}$$

So, $1 - \hat{\mathcal{D}}(\alpha\gamma)$ is the fraction of customers who will be eliminated while waiting in the queue. Of course, a (successful) customer can still be eliminated during his service time. The probability that a successful

customer will leave the system at the end of his service time S is

$$\begin{aligned}\mathbb{P}(E(\gamma) > S) &= \int_0^\infty \mathbb{P}(E(\gamma) > S \mid S = t) f_S(t) \mathrm{d}t = \hat{\mathcal{F}}_S(\gamma) \\ &= \left(1 - \frac{b\mu_1}{\mu_1 - \mu_2}\right) \frac{\mu_1}{\mu_1 + \gamma} + \frac{b\mu_1}{\mu_1 - \mu_2} \cdot \frac{\mu_2}{\mu_2 + \gamma}.\end{aligned}$$

We can conclude that a typical customer will leave the system at the end of his Coxian-2 service time is

$$\mathbb{P}(E(\alpha\gamma) > D)\mathbb{P}(E(\gamma) > S) = \hat{\mathcal{D}}(\alpha\gamma)\hat{\mathcal{F}}_S(\gamma).$$

Notice that all these results are justified by the memoryless property of the exponential distribution, and the independence of the two Poisson streams of local and complete catastrophes.

Finally, we return to the important question to calculate the *conditional* waiting-time distribution, *given* that a customer is successful, in fewer words, *the waiting-time distribution of a successful customer*,

$$\mathbb{P}(D > t | E(\alpha\gamma) > D) = \frac{\mathbb{P}(E(\alpha\gamma) > D > t)}{\mathbb{P}(E(\alpha\gamma) > D)} = \frac{1}{\hat{\mathcal{D}}(\alpha\gamma)} \int_t^\infty \int_y^\infty \alpha\gamma \mathrm{e}^{-\alpha\gamma x} f_D(y) \mathrm{d}x \mathrm{d}y =$$

$$\frac{1}{\hat{\mathcal{D}}(\alpha\gamma)} \int_t^\infty \mathrm{e}^{-\alpha\gamma y} f_D(y) \mathrm{d}y,$$

and with these tail-probabilities we can find the *conditional* expected waiting time of a customer in the queue, *given* that a customer is successful, or shortly, *the average waiting time of a successful customer*

$$\mathbb{E}[D | E(\alpha\gamma) > D] = \frac{1}{\hat{\mathcal{D}}(\alpha\gamma)} \int_0^\infty \int_t^\infty \mathrm{e}^{-\alpha\gamma y} f_D(y) \mathrm{d}y \mathrm{d}t = \frac{1}{\hat{\mathcal{D}}(\alpha\gamma)} \int_0^\infty \int_0^y \mathrm{e}^{-\alpha\gamma y} f_D(y) \mathrm{d}t \mathrm{d}y =$$

$$\frac{1}{\hat{\mathcal{D}}(\alpha\gamma)} \int_0^\infty y \mathrm{e}^{-\alpha\gamma y} f_D(y) \mathrm{d}y = -\frac{\hat{\mathcal{D}}'(\alpha\gamma)}{\hat{\mathcal{D}}(\alpha\gamma)}.$$

Next, we calculate the LST of the overall waiting time, say W_q, i.e. the waiting time in the queue of a typical customer (not necessarily successful). Then we have that $W_q = \min\{E(\alpha\gamma), D\}$, so the tail probabilities we easily find due to the independence of $E(\alpha\gamma)$ and D,

$$\mathbb{P}(W_q > t) = \mathrm{e}^{-\alpha\gamma t}\mathbb{P}(D > t).$$

Now recall that for a random variable $X \sim F_X(\cdot)$ we have the equality

$$\int_0^\infty \mathrm{e}^{-st}(1 - F_X(t))\mathrm{d}t = \frac{1}{s}\left(1 - \hat{\mathcal{F}}_X(s)\right).$$

So we get

$$\hat{\mathcal{W}}_q(s) = \frac{\alpha\gamma}{s + \alpha\gamma} + \frac{s}{s + \alpha\gamma}\hat{\mathcal{D}}(s + \alpha\gamma). \tag{17}$$

And now we find for the average overall waiting time

$$\overline{W}_q = -\hat{\mathcal{W}}_q'(0) = \frac{1}{\alpha\gamma}\left(1 - \hat{\mathcal{D}}(\alpha\gamma)\right). \tag{18}$$

Finally, comparing formula (13) for the average queue length L_q and formula (18) for the overall average waiting time in the queue $\overline{W}_q$, from Little's Law $L_q = \lambda\overline{W}_q$ we can conclude that

$$\hat{\mathcal{D}}(\alpha\gamma) = \frac{(1 - z_0)(\gamma + \mu_1)(\gamma + \mu_2)}{(\gamma + b\mu_1 + \mu_2)[\lambda(1 - z_0) + \alpha\gamma]},$$

which gives us a simple formula for the long-run fraction of successful customers.

3 The steady-state analysis of the model with repair times

In this section we extend the model of Section 2 with repair times, i.e. upon occurrence of a catastrophe at time t the server goes in repair, denoted by the server state $C(t) = 3$. The repair times follow an exponential distribution with rate ν. While the server is under repair customers continue to arrive at the system with Poisson arrival rate λ, and also catastrophes continue to arrive at the system with rate γ. Notice that while the server is under repair only complete catastrophes affect the state of the system: they eliminate all customers in the queue (there is no customer at the server), but we assume that they do not influence the repair process, so local catastrophes do not change the state of the system. Using the CTMC and definitions of the previous section we can write down the following balance equations,

$$\begin{aligned}
\lambda p_{00} &= (1-b)\mu_1 p_{10} + \mu_2 p_{20} + \nu p_{30} && (19)\\
(\lambda+\gamma+\mu_1)p_{10} &= \lambda p_{00} + (1-b)\mu_1 p_{11} + \mu_2 p_{21} + \nu p_{31} && (20)\\
(\lambda+\gamma+\mu_1)p_{1n} &= \lambda p_{1,n-1} + (1-b)\mu_1 p_{1,n+1} + \mu_2 p_{2,n+1} + \nu p_{3,n+1}, \quad n=1,2,\ldots && (21)\\
(\lambda+\gamma+\mu_2)p_{2n} &= \lambda p_{2,n-1} + b\mu_1 p_{1n}, \quad n=0,1,2,\ldots. && (22)\\
(\lambda+\alpha\gamma+\nu)p_{30} &= \alpha\gamma(1-p_{00}) + (1-\alpha)\gamma(p_{10}+p_{20}) && (23)\\
(\lambda+\alpha\gamma+\nu)p_{3n} &= \lambda p_{3,n-1} + (1-\alpha)\gamma(p_{1n}+p_{2n}), \quad n=1,2,\ldots. && (24)
\end{aligned}$$

Using standard methods we find from Eqs. (19)–(24) the following system of equations for the partial PGFs $P_1(z)$, $P_2(z)$, and $P_3(z)$.

$$\begin{aligned}
P_1(z) &= \frac{\lambda(z-1)p_{00} + \mu_2 P_2(z) + \nu P_3(z)}{z(\lambda(1-z)+\gamma+\mu_1) - (1-b)\mu_1}, && (25)\\
P_2(z) &= \frac{b\mu_1}{\lambda(1-z)+\gamma+\mu_2} P_1(z), && (26)\\
P_3(z) &= \frac{\alpha\gamma(1-p_{00}) + (1-\alpha)\gamma[P_1(z)+P_2(z)]}{\lambda(1-z)+\alpha\gamma+\nu}, && (27)
\end{aligned}$$

and from these equations we find through simple substitutions

$$P_1(z) = \frac{(\lambda(1-z)+\gamma+\mu_2)\left[\nu\alpha\gamma - (\nu\alpha\gamma + \lambda^2(1-z)^2 + \lambda(\alpha\gamma+\nu)(1-z))p_{00}\right]}{\begin{array}{c}[z(\lambda(1-z)+\gamma+\mu_1)-(1-b)\mu_1](\lambda(1-z)+\gamma+\mu_2)(\lambda(1-z)+\alpha\gamma+\nu)+\\ -b\mu_1\mu_2(\lambda(1-z)+\alpha\gamma+\nu) - \nu(1-\alpha)\gamma(\lambda(1-z)+\gamma+b\mu_1+\mu_2)\end{array}}. \quad (28)$$

Notice that for $\nu \to \infty$ expression (28) converges to expression (5) as expected.

To calculate the probability p_{00} for the model with repair times we have to show again that the denominator, now a *quartic* polynomial,

$$\begin{aligned}
\mathcal{D}(z) &= [z(\lambda(1-z)+\gamma+\mu_1)-(1-b)\mu_1](\lambda(1-z)+\gamma+\mu_2)(\lambda(1-z)+\alpha\gamma+\nu) +\\
&\quad -b\mu_1\mu_2(\lambda(1-z)+\alpha\gamma+\nu) - \nu(1-\alpha)\gamma(\lambda(1-z)+\gamma+b\mu_1+\mu_2) \qquad (29)
\end{aligned}$$

has a unique zero inside the unit disk $|z| < 1$. Because $\mathcal{D}(0) < 0$ and $\mathcal{D}(1) > 0$, there exists a real number $z_0 \in (0,1)$ for which $\mathcal{D}(z_0) = 0$ To show that z_0 is the only zero within the unit disk $|z| < 1$, we apply *Rouché's Theorem* again. Let

$$f(z) = [z(\lambda(1-z)+\gamma+\mu_1)-(1-b)\mu_1](\lambda(1-z)+\gamma+\mu_2)(\lambda(1-z)+\alpha\gamma+\nu),$$

and

$$g(z) = -b\mu_1\mu_2(\lambda(1-z)+\alpha\gamma+\nu) - \nu(1-\alpha)\gamma(\lambda(1-z)+\gamma+b\mu_1+\mu_2),$$

so $f(z)+g(z)=\mathcal{D}(z)$. Then $f(z)$ has exactly one zero within the unit disk $|z|<1$, because the first (quadratic) factor has two real zeroes, one in the interval $(0,1)$, and the other larger than 1, and the second and third factor have zeroes $z=(\lambda+\gamma+\mu_2)/\lambda>1$ and $z=(\lambda+\alpha\gamma+\nu)/\lambda>1$, respectively. Further, for all $z\in\Gamma=\{z\in\mathbb{C}:\ |z|=1\}$, using the triangle inequality we get the following upper bound for $|g(z)|$,

$$|g(z)|\leq b\mu_1\mu_2|\lambda(1-z)+\alpha\gamma+\nu|+\nu(1-\alpha)\gamma|\lambda(1-z)+\gamma+\mu_2|+\nu(1-\alpha)\gamma b\mu_1.$$

Next, for $z\in\Gamma$ we derive a lower bound for the absolute value of the first quadratic factor of $f(z)$. Using the zeroes, say $r_1\in(0,1)$ and $r_2>1$, we get for $z\in\Gamma$, using $|x-y|\geq||x|-|y||$ twice,

$$|z(\lambda(1-z)+\gamma+\mu_1)-(1-b)\mu_1|=|\lambda z^2-(\lambda+\gamma+\mu_1)z+(1-b)\mu_1|=\lambda|z-r_1|\cdot|z-r_2|$$

$$\geq\lambda||z|-r_1|\cdot|z-r_2|=\lambda(1-r_1)|z-r_2|\geq\lambda|(1-r_1)|z|-(1-r_1)r_2|=\lambda|1-(r_1+r_2)+r_1r_2|=$$

$$\lambda\left|1-\frac{\lambda+\gamma+\mu_1}{\lambda}+\frac{(1-b)\mu_1}{\lambda}\right|=\gamma+b\mu_1,$$

where we used the formulas for the sum and product of the roots of a quadratic equation. So we have also an (intermediate) lower bound for $|f(z)|$, namely

$$|f(z)|\geq(\gamma+b\mu_1)|\lambda(1-z)+\gamma+\mu_2|\cdot|\lambda(1-z)+\alpha\gamma+\nu|$$

Then using the upper and lower bound for $|g(z)|$ and $|f(z)|$, respectively, and using repeatedly $|x-y|\geq||x|-|y||\geq|x|-|y|$ we get for $z\in\Gamma$

$$\begin{aligned}\frac{|g(z)|}{|f(z)|}\ &\leq\ \frac{b\mu_1\mu_2|\lambda(1-z)+\alpha\gamma+\nu|+\nu(1-\alpha)\gamma|\lambda(1-z)+\gamma+\mu_2|+\nu(1-\alpha)\gamma b\mu_1}{|z(\lambda(1-z)+\gamma+\mu_1)-(1-b)\mu_1|\cdot|\lambda(1-z)+\gamma+\mu_2|\cdot|\lambda(1-z)+\alpha\gamma+\nu|}\leq\frac{1}{\gamma+b\mu_1}\\
&\times\left(\frac{b\mu_1\mu_2}{|\lambda(1-z)+\gamma+\mu_2|}+\frac{\nu(1-\alpha)\gamma}{|\lambda(1-z)+\alpha\gamma+\nu|}+\frac{\nu(1-\alpha)\gamma b\mu_1}{|\lambda(1-z)+\gamma+\mu_2|\cdot|\lambda(1-z)+\alpha\gamma+\nu|}\right)\leq\\
&\frac{1}{\gamma+b\mu_1}\left(\frac{b\mu_1\mu_2}{|\lambda+\gamma+\mu_2-\lambda|z||}+\frac{\nu(1-\alpha)\gamma}{|\lambda+\alpha\gamma+\nu-\lambda|z||}+\frac{\nu(1-\alpha)\gamma b\mu_1}{|\lambda+\gamma+\mu_2-\lambda|z||\cdot|\lambda+\alpha\gamma+\nu-\lambda|z||}\right)=\\
&\frac{1}{\gamma+b\mu_1}\left(\frac{b\mu_1\mu_2}{\gamma+\mu_2}+\frac{\nu(1-\alpha)\gamma}{\alpha\gamma+\nu}+\frac{\nu(1-\alpha)\gamma b\mu_1}{(\gamma+\mu_2)(\alpha\gamma+\nu)}\right)=\frac{b\mu_1\mu_2(\alpha\gamma+\nu)+\nu(1-\alpha)\gamma(\gamma+b\mu_1+\mu_2)}{(\gamma+b\mu_1)(\gamma+\mu_2)(\alpha\gamma+\nu)}<1.\end{aligned}$$

So, from Rouché's Theorem we can conclude that $\mathcal{D}(z)=f(z)+g(z)$ has a unique zero $z=z_0$ within the unit disk $|z|<1$. Hence, for $z=z_0$ also the numerator of expression (28) must vanish. So, for the model with repair times we find

$$p_{00}=\frac{\nu\alpha\gamma}{\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma}. \tag{30}$$

Notice that for $\nu\to\infty$, the probability p_{00} converges to the expression for p_{00} of the simplified model without repair times (see Eq. (7)), as expected.

3.1 The Queue Length Distribution of the model with repair times

Using *Ludovico Ferrari*'s general solution for a quartic equation, the four zeroes of the denominator $\mathcal{D}(z)$ in Eq. (29) can be obtained explicitly. Denoting these zeroes by z_0,z_1,z_2,z_3, we can write

$$\mathcal{D}(z)=-\lambda^3(z-z_0)(z-z_1)(z-z_2)(z-z_3),$$

where $z_0 \in (0,1)$ is the unique zero inside the unit disk and $|z_i| > 1$ for $i = 1, 2, 3$. Cancelling the factor $z - z_0$ in the expression for $P_1(z)$ given in Eq. (28) we get, using the expression for p_{00} given in Eq. (30),

$$P_1(z) = \frac{\nu\alpha\gamma(\lambda(1-z)+\gamma+\mu_2)[\lambda(2-z_0-z)+\alpha\gamma+\nu]}{\lambda^2[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma](z_1-z)(z_2-z)(z_3-z)}, \tag{31}$$

$$P_2(z) = \frac{b\mu_1\nu\alpha\gamma[\lambda(2-z_0-z)+\alpha\gamma+\nu]}{\lambda^2[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma](z_1-z)(z_2-z)(z_3-z)}, \tag{32}$$

$$\begin{aligned} P_3(z) &= \frac{\alpha\gamma[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)]}{[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma](\lambda(1-z)+\alpha\gamma+\nu)} \\ &+\frac{\nu\alpha(1-\alpha)\gamma^2(\lambda(1-z)+\gamma+b\mu_1+\mu_2)[\lambda(2-z_0-z)+\alpha\gamma+\nu]}{\lambda^2[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma](z_1-z)(z_2-z)(z_3-z)(\lambda(1-z)+\alpha\gamma+\nu)}. \end{aligned} \tag{33}$$

To calculate the steady-state distribution of the system-state (C, Q) we firstly focus on $P_2(z)$. After some simple algebra we find from Eq. (32)

$$\begin{aligned} P_2(z) &= \frac{b\mu_1\nu\alpha\gamma}{\lambda^2[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]} \times \left(\frac{\lambda(2-z_0-z_1)+\alpha\gamma+\nu}{(z_2-z_1)(z_3-z_1)} \cdot \frac{1}{z_1-z} \right. \\ &\left. + \frac{\lambda(2-z_0-z_2)+\alpha\gamma+\nu}{(z_1-z_2)(z_3-z_2)} \cdot \frac{1}{z_2-z} + \frac{\lambda(2-z_0-z_3)+\alpha\gamma+\nu}{(z_1-z_3)(z_2-z_3)} \cdot \frac{1}{z_3-z} \right), \end{aligned} \tag{34}$$

and hence we find for $n = 0, 1, 2, \ldots$

$$\begin{aligned} p_{2n} &= \frac{b\mu_1\nu\alpha\gamma}{\lambda^2[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]} \times \left[\frac{\lambda(2-z_0-z_1)+\alpha\gamma+\nu}{z_1(z_2-z_1)(z_3-z_1)} \left(\frac{1}{z_1}\right)^n \right. \\ &\left. + \frac{\lambda(2-z_0-z_2)+\alpha\gamma+\nu}{z_2(z_1-z_2)(z_3-z_2)} \left(\frac{1}{z_2}\right)^n + \frac{\lambda(2-z_0-z_3)+\alpha\gamma+\nu}{z_3(z_1-z_3)(z_2-z_3)} \left(\frac{1}{z_3}\right)^n \right]. \end{aligned} \tag{35}$$

As in Section 2 we have again

$$p_{1n} = \frac{\lambda+\gamma+\mu_2}{b\mu_1} p_{2n} - \frac{\lambda}{b\mu_1} p_{2,n-1}, \quad n = 0, 1, 2, \ldots$$

where we define $p_{2,-1} = 0$.

To calculate the probabilities p_{3n} we start from the expression for $P_3(z)$ given in Eq. (33) and denote the two terms of (33) as $R(z)$ and $T(z)$, respectively. Then the first term $R(z)$ can be easily rewritten as a power series

$$R(z) = \frac{\alpha\gamma[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)]}{(\lambda+\alpha\gamma+\nu)[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]} \sum_{n=0}^{\infty} \left(\frac{\lambda}{\lambda+\alpha\gamma+\nu}\right)^n z^n.$$

To write the second term $T(z)$ as a (sum of) power series we rewrite the second factor of $T(z)$ as

$$T(z) = \frac{\nu\alpha(1-\alpha)\gamma^2}{\lambda^3[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]} \left[\frac{c_1}{z_1-z} + \frac{c_2}{z_2-z} + \frac{c_3}{z_3-z} + \frac{c_4}{\frac{\lambda+\alpha\gamma+\nu}{\lambda}-z} \right] =$$

$$\frac{\nu\alpha(1-\alpha)\gamma^2}{\lambda^3[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]} \left[\frac{c_1}{z_1} \sum_{n=0}^{\infty} \left(\frac{1}{z_1}\right)^n z^n + \frac{c_2}{z_2} \sum_{n=0}^{\infty} \left(\frac{1}{z_2}\right)^n z^n + \frac{c_3}{z_3} \sum_{n=0}^{\infty} \left(\frac{1}{z_3}\right)^n z^n \right]$$

$$+\frac{\nu\alpha(1-\alpha)\gamma^2}{\lambda^3[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]} \cdot \frac{\lambda c_4}{\lambda+\alpha\gamma+\nu} \sum_{n=0}^{\infty} \left(\frac{\lambda}{\lambda+\alpha\gamma+\nu}\right)^n z^n.$$

The coefficients c_1, c_2, c_3, and c_4 are obtained by a standard partial fraction expansion, where

$$z_4 := \frac{\lambda + \alpha\gamma + \nu}{\lambda}.$$

The resulting expressions are

$$\begin{aligned}
c_1 &= \frac{\lambda(\lambda(1-z_1)+\gamma+b\mu_1+\mu_2)(\lambda(2-z_0-z_1)+\alpha\gamma+\nu)}{(z_2-z_1)(z_3-z_1)(\lambda(1-z_1)+\alpha\gamma+\nu)} \\
c_2 &= \frac{\lambda(\lambda(1-z_2)+\gamma+b\mu_1+\mu_2)(\lambda(2-z_0-z_2)+\alpha\gamma+\nu)}{(z_1-z_2)(z_3-z_2)(\lambda(1-z_2)+\alpha\gamma+\nu)} \\
c_3 &= \frac{\lambda(\lambda(1-z_3)+\gamma+b\mu_1+\mu_2)(\lambda(2-z_0-z_3)+\alpha\gamma+\nu)}{(z_1-z_3)(z_2-z_3)(\lambda(1-z_3)+\alpha\gamma+\nu)} \\
c_4 &= \frac{(\lambda(1-z_4)+\gamma+b\mu_1+\mu_2)(\lambda(2-z_0-z_4)+\alpha\gamma+\nu)}{(z_1-z_4)(z_2-z_4)(z_3-z_4)}.
\end{aligned}$$

So, the expressions for the probabilities p_{3n} are given as follows ($n = 0, 1, 2, \ldots$)

$$\begin{aligned}
p_{3n} &= \left[\frac{\alpha\gamma[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)]}{(\lambda+\alpha\gamma+\nu)[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]}\right]\left(\frac{\lambda}{\lambda+\alpha\gamma+\nu}\right)^n \\
&\quad + \left[\frac{\nu\alpha(1-\alpha)\gamma^2\cdot c_4}{\lambda^2[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma](\lambda+\alpha\gamma+\nu)}\right]\left(\frac{\lambda}{\lambda+\alpha\gamma+\nu}\right)^n \\
&\quad + \frac{\nu\alpha(1-\alpha)\gamma^2}{\lambda^3[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]}\left[c_1\left(\frac{1}{z_1}\right)^{n+1}+c_2\left(\frac{1}{z_2}\right)^{n+1}+c_3\left(\frac{1}{z_3}\right)^{n+1}\right] = \\
&\quad \frac{\nu\alpha(1-\alpha)\gamma^2}{\lambda^3[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]}\left[c_1\left(\frac{1}{z_1}\right)^{n+1}+c_2\left(\frac{1}{z_2}\right)^{n+1}+c_3\left(\frac{1}{z_3}\right)^{n+1}\right]. \quad (36)
\end{aligned}$$

The last step follows from

$$\frac{\alpha\gamma[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)]}{(\lambda+\alpha\gamma+\nu)[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]}+\frac{\nu\alpha(1-\alpha)\gamma^2\cdot c_4}{\lambda^2[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma](\lambda+\alpha\gamma+\nu)}=0,$$

which can be seen after some algebra using the expression for c_4 and the observation that

$$(z_1-z_4)(z_2-z_4)(z_3-z_4)=\frac{\mathcal{D}(z_4)}{\lambda^3(z_4-z_0)}=\frac{-\nu(1-\alpha)\gamma[(1-\alpha)\gamma+b\mu_1+\mu_2]}{\lambda^2[\lambda(1-z_0)+\alpha\gamma+\nu]}.$$

Once we have calculated the p_{1n}, p_{2n}, and p_{3n} we can calculate the queue length probabilities $q_n := \mathbb{P}(Q=n)$. We get from Eqs.(35) and (36) for $n = 0, 1, 2, \ldots$

$$
\begin{aligned}
q_n &= p_{00}\mathbf{I}_{\{n=0\}} + p_{1n} + p_{2n} + p_{3n} = p_{00}\mathbf{I}_{\{n=0\}} + \left(\frac{\lambda+\gamma+\mu_2}{b\mu_1}+1\right)p_{2n} - \frac{\lambda}{b\mu_1}p_{2,n-1} + p_{3n} = \\
&\frac{\nu\alpha\gamma}{\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma}\mathbf{I}_{\{n=0\}} \\
&+\frac{\nu\alpha\gamma}{\lambda^2[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]} \\
&\times\left[\frac{(\lambda(2-z_0-z_1)+\alpha\gamma+\nu)(\lambda(1-z_1)+\gamma+b\mu_1+\mu_2)}{z_1(z_2-z_1)(z_3-z_1)}\left(\frac{1}{z_1}\right)^n\right. \\
&+\frac{(\lambda(2-z_0-z_2)+\alpha\gamma+\nu)(\lambda(1-z_2)+\gamma+b\mu_1+\mu_2)}{z_2(z_1-z_2)(z_3-z_2)}\left(\frac{1}{z_2}\right)^n \\
&\left.+\frac{(\lambda(2-z_0-z_3)+\alpha\gamma+\nu)(\lambda(1-z_3)+\gamma+b\mu_1+\mu_2)}{z_3(z_1-z_3)(z_2-z_3)}\left(\frac{1}{z_3}\right)^n\right] \\
&+\frac{\nu\alpha(1-\alpha)\gamma^2}{\lambda^3[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]}\left[c_1\left(\frac{1}{z_1}\right)^{n+1}+c_2\left(\frac{1}{z_2}\right)^{n+1}+c_3\left(\frac{1}{z_3}\right)^{n+1}\right] \quad (37)
\end{aligned}
$$

Expression (37) can be simplified further by summing the coefficients, using the expressions for c_1, c_2, c_3. After some simple algebra we get

$$
\begin{aligned}
q_n &= \frac{\nu\alpha\gamma}{\lambda^2\left[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma\right]}\left[\lambda^2\mathbf{I}_{\{n=0\}}\right. \\
&+\frac{(\lambda(2-z_0-z_1)+\alpha\gamma+\nu)(\lambda(1-z_1)+\gamma+b\mu_1+\mu_2)(\lambda(1-z_1)+\gamma+\nu)}{z_1(z_2-z_1)(z_3-z_1)(\lambda(1-z_1)+\alpha\gamma+\nu)}\left(\frac{1}{z_1}\right)^n \\
&+\frac{(\lambda(2-z_0-z_2)+\alpha\gamma+\nu)(\lambda(1-z_2)+\gamma+b\mu_1+\mu_2)(\lambda(1-z_2)+\gamma+\nu)}{z_2(z_1-z_2)(z_3-z_2)(\lambda(1-z_2)+\alpha\gamma+\nu)}\left(\frac{1}{z_2}\right)^n \\
&\left.+\frac{(\lambda(2-z_0-z_3)+\alpha\gamma+\nu)(\lambda(1-z_3)+\gamma+b\mu_1+\mu_2)(\lambda(1-z_3)+\gamma+\nu)}{z_3(z_1-z_3)(z_2-z_3)(\lambda(1-z_3)+\alpha\gamma+\nu)}\left(\frac{1}{z_3}\right)^n\right]. \quad (38)
\end{aligned}
$$

Again using the normalization condition and Eqs. (26) and (27), the long-run fractions of time that the server is in the different states are

$$
\begin{aligned}
\mathbb{P}(C=0) &= p_{00}, \\
\mathbb{P}(C=1) &= P_1(1) = \frac{\nu(\gamma+\mu_2)}{(\gamma+b\mu_1+\mu_2)(\gamma+\nu)}(1-p_{00}), \\
\mathbb{P}(C=2) &= P_2(1) = \frac{b\mu_1\nu}{(\gamma+b\mu_1+\mu_2)(\gamma+\nu)}(1-p_{00}), \\
\mathbb{P}(C=3) &= P_3(1) = \frac{\gamma}{\gamma+\nu}(1-p_{00}).
\end{aligned}
$$

As in Section 2 we can calculate the long-run average number of customers in the queue, again denoted as L_q, very easily using the above expression for $P_2(z)$. We find from Eqs. (34), (26), and

$$
\begin{aligned}
P_2'(z) &= \frac{b\mu_1\nu\alpha\gamma}{\lambda^2[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]}\times\left(\frac{\lambda(2-z_0-z_1)+\alpha\gamma+\nu}{(z_2-z_1)(z_3-z_1)}\cdot\frac{1}{(z_1-z)^2}\right. \\
&\left.+\frac{\lambda(2-z_0-z_2)+\alpha\gamma+\nu}{(z_1-z_2)(z_3-z_2)}\cdot\frac{1}{(z_2-z)^2}+\frac{\lambda(2-z_0-z_3)+\alpha\gamma+\nu}{(z_1-z_3)(z_2-z_3)}\cdot\frac{1}{(z_3-z)^2}\right), \quad (39) \\
P_1'(z) &= \frac{\lambda(1-z)+\gamma+\mu_2}{b\mu_1}P_2'(z)-\frac{\lambda}{b\mu_1}P_2(z). \quad (40)
\end{aligned}
$$

To calculate the derivative of $P_3(z)$ we rewrite Eq. (27) as

$$P_3(z) = \frac{\alpha\gamma(1-p_{00})}{\lambda} \cdot \frac{1}{\frac{\lambda+\alpha\gamma+\nu}{\lambda} - z} + \frac{(1-\alpha)\gamma}{b\mu_1} \cdot \frac{\lambda(1-z)+\gamma+b\mu_1+\mu_2}{\lambda(1-z)+\alpha\gamma+\nu} \cdot P_2(z). \tag{41}$$

Differentiating $P_3(z)$ using expression (41) gives

$$P_3'(z) = \frac{\alpha\gamma(1-p_{00})}{\lambda\left(\frac{\lambda+\alpha\gamma+\nu}{\lambda} - z\right)^2} + \frac{(1-\alpha)\gamma}{b\mu_1}\left[\frac{\lambda[(1-\alpha)\gamma+b\mu_1+\mu_2-\nu]}{[\lambda(1-z)+\alpha\gamma+\nu]^2}P_2(z) + \frac{\lambda(1-z)+\gamma+b\mu_1+\mu_2}{\lambda(1-z)+\alpha\gamma+\nu}P_2'(z)\right]. \tag{42}$$

With the above results, after some simplification, we find

$$\begin{aligned} L_q &= \sum_{n=1}^{\infty} n(p_{1n}+p_{2n}+p_{3n}) = P_1'(1)+P_2'(1)+P_3'(1) \\ &= \left[\frac{\alpha\gamma\lambda}{(\alpha\gamma+\nu)^2} + \frac{\lambda\nu\left[(1-\alpha)\gamma\left((1-\alpha)\gamma+b\mu_1+\mu_2-\nu\right)-(\alpha\gamma+\nu)^2\right]}{(\alpha\gamma+\nu)^2(\gamma+b\mu_1+\mu_2)(\gamma+\nu)}\right](1-p_{00}) \\ &\quad + \frac{(\gamma+\nu)(\gamma+b\mu_1+\mu_2)}{b\mu_1(\alpha\gamma+\nu)}P_2'(1), \end{aligned} \tag{43}$$

where $P_2'(1)$ can be calculated directly from Eq. (39). To obtain a more convenient expression for $P_2'(1)$, we return to the unfactored form of $P_2(z)$. Combining Eqs. (26) and (28), and using the denominator $\mathcal{D}(z)$ defined in Eq. (29), gives

$$P_2(z) = \frac{b\mu_1\left[\nu\alpha\gamma - (\nu\alpha\gamma+\lambda^2(1-z)^2+\lambda(\alpha\gamma+\nu)(1-z))\,p_{00}\right]}{\mathcal{D}(z)}. \tag{44}$$

Direct differentiation at $z=1$, followed by some tedious algebra, gives

$$P_2'(1) = \frac{b\mu_1}{\alpha\gamma} \cdot \frac{\lambda(\gamma+b\mu_1+\mu_2)(\gamma+\nu)(\alpha\gamma+\nu) - \begin{bmatrix}\nu(\gamma+\mu_1)(\gamma+\mu_2)(\alpha\gamma+\nu)\\ -\lambda\nu\alpha\gamma(\gamma+\nu)\\ +\lambda\alpha\gamma^2(\gamma+b\mu_1+\mu_2)\end{bmatrix}(1-p_{00})}{[(\gamma+b\mu_1+\mu_2)(\gamma+\nu)]^2}. \tag{45}$$

Now, plugging in expression (45) in formula (43) we get after some tedious algebra, and using Eq. (30)

$$\begin{aligned} L_q &= \frac{1}{\alpha\gamma}\left[\lambda - \frac{\nu(\gamma+\mu_1)(\gamma+\mu_2)}{(\gamma+b\mu_1+\mu_2)(\gamma+\nu)}(1-p_{00})\right] = \\ &\quad \frac{\lambda}{\alpha\gamma}\left[1 - \frac{\nu(\gamma+\mu_1)(\gamma+\mu_2)[\lambda(1-z_0)^2+(\alpha\gamma+\nu)(1-z_0)]}{(\gamma+b\mu_1+\mu_2)(\gamma+\nu)[\lambda^2(1-z_0)^2+\lambda(\alpha\gamma+\nu)(1-z_0)+\nu\alpha\gamma]}\right], \end{aligned} \tag{46}$$

and again we see that for $\nu\to\infty$ this expression converges to formula (13) for the simple model.

3.2 The Waiting Time Distribution(s) for the model with repair times

Next, we will calculate the waiting-time distribution of a successful customer (see paragraph 2.2 for the definition) for the model with repair times. As in Section 2 we introduce the random variable D for the waiting time of a typical customer as if complete catastrophes are temporarily neglected, and calculate the probability distribution of D by conditioning on the system-state (C,Q) seen by a typical customer

upon his arrival. Again as in Section 2 we call D the *virtual waiting time* of a typical customer, i.e. the amount of work present in the system at the time of arrival (the workload), in other words the waiting time 'foreseen' upon arrival by a typical customer *before he might be eliminated from the system by the occurrence of a complete catastrophe*. Let R be a generic repair time, so $R \sim \text{Exp}(\nu)$. Further, we define a so-called 'single service/repair time'

$$T^* := \min\{E((1-\alpha)\gamma), S\} + R\mathbf{I}_{\{E((1-\alpha)\gamma)<S\}}.$$

So, T^* is the time lapse from the start of the service of a customer until the server is idle again or ready for the service of the next customer. Now, let T_j^* be the single service/repair time of the j-th customer 'observed' by the typical customer, enumerated from the moment of his arrival, and R_j the repair time with which this j-th customer *might be confronted* in case this customer will be eliminated due to the arrival of a (local) catastrophe. Then we can write down the following conditional probabilities for the tail probabilities of the virtual waiting time D.

$$\begin{aligned}
\mathbb{P}(D > t \mid C = 1; Q = n) &= \mathbb{P}\left(\sum_{j=1}^{n+1} T_j^* > t\right) \\
\mathbb{P}(D > t \mid C = 2; Q = n) &= \mathbb{P}\left(\min\{E((1-\alpha)\gamma), E(\mu_2)\} + R_1\mathbf{I}_{\{E((1-\alpha)\gamma)<E(\mu_2)\}} + \sum_{j=2}^{n+1} T_j^* > t\right) \\
\mathbb{P}(D > t \mid C = 3; Q = n) &= \mathbb{P}\left(R_1 + \sum_{j=2}^{n+1} T_j^* > t\right).
\end{aligned}$$

To evaluate these conditional probabilities we first calculate the tail probabilities of the random variable T^*. Conditioning on whether a local catastrophe or service completion occurs first, and using the independence and the respective distributions of the service, catastrophe, and repair times, we obtain

$$\begin{aligned}
\mathbb{P}(T^* > t) &= \left(1 - \frac{b\mu_1}{\mu_1 - \mu_2}\right) \frac{\mu_1 - \nu}{\mu_1 + (1-\alpha)\gamma - \nu} \mathrm{e}^{-(\mu_1 + (1-\alpha)\gamma)t} + \frac{b\mu_1}{\mu_1 - \mu_2} \cdot \frac{\mu_2 - \nu}{\mu_2 + (1-\alpha)\gamma - \nu} \mathrm{e}^{-(\mu_2 + (1-\alpha)\gamma)t} \\
&\quad + \left[\left(1 - \frac{b\mu_1}{\mu_1 - \mu_2}\right) \frac{(1-\alpha)\gamma}{\mu_1 + (1-\alpha)\gamma - \nu} + \frac{b\mu_1}{\mu_1 - \mu_2} \cdot \frac{(1-\alpha)\gamma}{\mu_2 + (1-\alpha)\gamma - \nu}\right] \mathrm{e}^{-\nu t}, \qquad (47)
\end{aligned}$$

and from the tail probabilities we get the LST of T^*

$$\begin{aligned}
\hat{\mathcal{F}}_{T^*}(s) &= \left(1 - \frac{b\mu_1}{\mu_1 - \mu_2}\right) \frac{\mu_1 - \nu}{\mu_1 + (1-\alpha)\gamma - \nu} \cdot \frac{\mu_1 + (1-\alpha)\gamma}{s + \mu_1 + (1-\alpha)\gamma} \\
&\quad + \frac{b\mu_1}{\mu_1 - \mu_2} \cdot \frac{\mu_2 - \nu}{\mu_2 + (1-\alpha)\gamma - \nu} \cdot \frac{\mu_2 + (1-\alpha)\gamma}{s + \mu_2 + (1-\alpha)\gamma} \\
&\quad + \left[\left(1 - \frac{b\mu_1}{\mu_1 - \mu_2}\right) \frac{(1-\alpha)\gamma}{\mu_1 + (1-\alpha)\gamma - \nu} + \frac{b\mu_1}{\mu_1 - \mu_2} \cdot \frac{(1-\alpha)\gamma}{\mu_2 + (1-\alpha)\gamma - \nu}\right] \frac{\nu}{s + \nu}. \qquad (48)
\end{aligned}$$

To calculate the LST of D, say again $\hat{\mathcal{D}}(s)$, we also need the LST of the random variable

$$U^* := \min\{E((1-\alpha)\gamma), E(\mu_2)\} + R\mathbf{I}_{\{E((1-\alpha)\gamma)<E(\mu_2)\}},$$

which represents the remaining service/repair time when the customer in service is in the second service phase. Its distribution is obtained in the same way as that of T^*. Conditioning on whether a local catastrophe or completion of the second service phase occurs first, we obtain

$$\mathbb{P}(U^* > u) = \frac{\mu_2 - \nu}{\mu_2 + (1-\alpha)\gamma - \nu} \mathrm{e}^{-(\mu_2 + (1-\alpha)\gamma)u} + \frac{(1-\alpha)\gamma}{\mu_2 + (1-\alpha)\gamma - \nu} \mathrm{e}^{-\nu u}.$$

Hence, the LST of U^* is

$$\hat{\mathcal{F}}_{U^*}(s) = \frac{\mu_2 - \nu}{\mu_2 + (1-\alpha)\gamma - \nu} \cdot \frac{\mu_2 + (1-\alpha)\gamma}{s + \mu_2 + (1-\alpha)\gamma} + \frac{(1-\alpha)\gamma}{\mu_2 + (1-\alpha)\gamma - \nu} \cdot \frac{\nu}{s+\nu}.$$

Combining all these previous results we find for the LST of the virtual waiting time D, again using the *Law of Total Probability* and the PASTA property

$$\hat{\mathcal{D}}(s) = p_{00} + \hat{\mathcal{F}}_{T^*}(s) P_1 \left(\hat{\mathcal{F}}_{T^*}(s)\right) + \hat{\mathcal{F}}_{U^*}(s) P_2 \left(\hat{\mathcal{F}}_{T^*}(s)\right) + \frac{\nu}{s+\nu} P_3 \left(\hat{\mathcal{F}}_{T^*}(s)\right). \tag{49}$$

Now we have reached the point that we can calculate the waiting-time distribution of a successful customer. Firstly, using D as in Section 2, we can give the probability that a typical customer is successful, which is simply the probability that during his virtual waiting time no complete catastrophe occurs, so

$$\mathbb{P}(\text{customer is successful}) = \mathbb{P}(E(\alpha\gamma) > D) = \hat{\mathcal{D}}(\alpha\gamma).$$

Hence the probability that a customer *leaves* the system successfully, i.e. after being served can be calculated as follows

$$\mathbb{P}(\text{customer leaves the system after being served}) = \mathbb{P}(E(\alpha\gamma) > D)\mathbb{P}(E(\gamma) > S) = \hat{\mathcal{D}}(\alpha\gamma)\hat{\mathcal{F}}_S(\gamma).$$

Finally, *verbatim* following the reasoning in Section 2, we calculate the *conditional* waiting-time distribution, *given* that a customer is successful, in fewer words, *the waiting-time distribution of a successful customer*, i.e.

$$\mathbb{P}(D > t \mid E(\alpha\gamma) > D) = \frac{1}{\hat{\mathcal{D}}(\alpha\gamma)} \int_t^\infty \mathrm{e}^{-\alpha\gamma y} f_D(y) \mathrm{d}y,$$

and with these tail-probabilities we can find the *expected waiting time of a successful customer*,

$$\mathbb{E}[D \mid E(\alpha\gamma) > D] = -\frac{\hat{\mathcal{D}}'(\alpha\gamma)}{\hat{\mathcal{D}}(\alpha\gamma)}.$$

We end also this section with the overall waiting time of a typical customer, say again W_q. Following *verbatim* the reasoning presented at the end of Section 2 we find again

$$\hat{\mathcal{W}}_q(s) = \frac{\alpha\gamma}{s+\alpha\gamma} + \frac{s}{s+\alpha\gamma}\hat{\mathcal{D}}(s+\alpha\gamma). \tag{50}$$

And hence for the average overall waiting time

$$\overline{W}_q = -\hat{\mathcal{W}}_q'(0) = \frac{1}{\alpha\gamma}\left(1 - \hat{\mathcal{D}}(\alpha\gamma)\right). \tag{51}$$

And again we find a simple formula for the long-run fraction of successful customers, i.e. $\hat{\mathcal{D}}(\alpha\gamma)$, this time from Eq. (46). Using Little's Law $L_q = \lambda\overline{W}_q$, we get for the model with repair times

$$\hat{\mathcal{D}}(\alpha\gamma) = \frac{\nu(\gamma+\mu_1)(\gamma+\mu_2)[\lambda(1-z_0)^2 + (\alpha\gamma+\nu)(1-z_0)]}{(\gamma + b\mu_1 + \mu_2)(\gamma+\nu)[\lambda^2(1-z_0)^2 + \lambda(\alpha\gamma+\nu)(1-z_0) + \nu\alpha\gamma]}.$$

4 Numerical results

Once we have derived the formulas for the in our opinion most relevant performance measures of the two models with Bernoulli catastrophes, in this section we will illustrate the phenomenon of non-monotonicity announced at the end of Section 1 by presenting several numerical results which will be contrasted with analogous results for what we will call 'auxiliary' $M/G/1$ models. In all our numerical examples we use the so-called *gamma-normalisation* for the Coxian-2 distributed service time S, i.e. for a given mean $\mathbb{E}[S]$ and squared coefficient of variation $c_S^2 = \text{VAR}(S)/\mathbb{E}^2[S] > 1/2$ (we exclude $c_S^2 = 1/2$, because then $\mu_1 = \mu_2 =: \mu$, making S Erlang$(2,\mu)$ distributed, which we did not incorporate in our steady-state analysis when we studied the waiting-time distribution) we choose the following two-moment match:

$$\mu_1 = \frac{2}{\mathbb{E}[S]}\left(1 + \sqrt{\frac{c_S^2 - \frac{1}{2}}{c_S^2 + 1}}\right), \quad \mu_2 = \frac{4}{\mathbb{E}[S]} - \mu_1, \quad b = \frac{\mu_2}{\mu_1}(\mu_1 \mathbb{E}[S] - 1)$$

which has the nice property that the third moment $\mathbb{E}[S^3]$ is equal to the third moment of a gamma-distributed random variable with the same first and second moments as S (see TIJMS [20] for more details). This choice is convenient for illustrating the sensitivity of the performance measures to the (ir)regularity of the service time: we vary c_S^2 while keeping $\mathbb{E}[S]$ fixed, and hence the offered load $\varrho := \lambda \mathbb{E}[S]$, constant. Notice that, due to the complete catastrophes, stability does not require $\varrho < 1$. In the examples below we take

$$\mathbb{E}[S] = 1, \qquad \varrho = 0.90, \qquad \gamma = 0.05, \qquad \alpha = 0.5,$$

so $\lambda = 0.90$. For the model with repair times we take $\nu = 0.25$.

From the perspective of the classical $M/G/1$ queue, the behaviour observed below is perhaps surprising. If $\mathbb{E}[S]$ and hence the offered load $\varrho = \lambda \mathbb{E}[S] < 1$ are kept fixed, the Pollaczek–Khinchine formula gives

$$L_q^{\text{P–K}}(S) = \frac{\varrho^2(1 + c_S^2)}{2(1 - \varrho)}. \tag{52}$$

Consequently, both the mean queue length and, by Little's Law, the mean waiting time are strictly increasing and unbounded in c_S^2. Moreover, in the classical model every customer eventually reaches the server and completes service. One might therefore expect increasing service-time irregularity to have an increasingly adverse effect on the performance of our models as well.

Our numerical results, presented in Tables 1 and 2, show that the dependence on c_S^2 is non-monotone. In the models without and with repair times, the mean queue length, the mean virtual waiting time, and the mean waiting time of a successful customer first increase and subsequently decrease with an increasing c_S^2. Correspondingly, the fractions of successful and fully successful customers first decrease and subsequently increase with an increasing c_S^2. Thus, the classical Pollaczek–Khinchine intuition is valid for small and moderate values of c_S^2, but greater service-time irregularity eventually improves the performance measures considered here.

Performance measures. We first compare the mean queue length L_q of the models without and with repair times as c_S^2 varies, *ceteris paribus*. We then consider the mean virtual waiting time of a typical arrival, the mean waiting time of a successful customer, and the fraction of *fully successful* customers, i.e. customers who leave the system after completing service.

For the model without repair times, the mean queue length L_q is given by (13), and the LST $\hat{\mathcal{D}}(s)$ of the virtual waiting time by (16). For the model with repair times, the corresponding expressions

are (46) and (49). Recall that D is the time until a tagged customer reaches the server if complete catastrophes are temporarily neglected after his arrival. The realised waiting time of an arbitrary customer is $W_q = \min\{E(\alpha\gamma), D\}$. By Little's Law,

$$L_q = \lambda \overline{W}_q = \lambda \mathbf{E}[\min\{D, E(\alpha\gamma)\}], \tag{53}$$

meaning that each arriving customer is counted in the queue for his virtual waiting time, cut off at the first complete catastrophe after his arrival. We recall that from the two (different!) explicit expressions for $\hat{\mathcal{D}}(s)$ derived in Sections 2 and 3 we have for both models

$$\mathbf{E}[D] = -\hat{\mathcal{D}}'(0), \qquad \mathbf{E}[D \mid \mathrm{s}] = -\frac{\hat{\mathcal{D}}'(\alpha\gamma)}{\hat{\mathcal{D}}(\alpha\gamma)}, \qquad \mathbf{P}(\mathrm{s}) = \hat{\mathcal{D}}(\alpha\gamma),$$

where s denotes the event that the customer reaches the server. Conditional on this event, service is completed with probability $\hat{\mathcal{F}}_S(\gamma)$. It follows that

$$\mathbf{P}(\text{fully s.}) = \hat{\mathcal{D}}(\alpha\gamma)\hat{\mathcal{F}}_S(\gamma)$$

is the fraction of arrivals that leave the system after completing service. The transform value $\hat{\mathcal{F}}_S(\gamma)$ will also be used below to quantify the mean duration of an active service attempt and the probability that a repair is required.

Auxiliary comparison queues. To distinguish the effects of local catastrophes, complete catastrophes, and repair times, we compare the full models with two auxiliary $M/G/1$ queues in which complete catastrophes are suspended. These auxiliary queues retain the local catastrophes, which interrupt the customer in service without clearing the queue. Their generic service times are the random variables S^* and T^* introduced in Sections 2 and 3. For clarity, we repeat their definitions. Write

$$\theta := (1-\alpha)\gamma,$$

and let the generic random variable $E(\theta)$ be exponentially distributed with rate θ. Without repair times, the generic service time of the auxiliary queue is

$$S^* := \min\{S, E(\theta)\}.$$

With repair times, let R be exponentially distributed with rate ν and independent of S and $E(\theta)$, and define

$$T^* := S^* + R\,\mathbf{1}_{\{E(\theta)<S\}}.$$

Thus, T^* includes a repair time precisely when a local catastrophe occurs before service completion.

The four comparison models can be summarised as follows:

	Without repair times	With repair times
Local catastrophes only	$M/G/1$ with S^*	$M/G/1$ with T^*
Local and complete catastrophes	Full model without repairs	Full model with repairs

For each auxiliary queue, let $X \in \{S^*, T^*\}$ denote its service time. Write $\varrho_X := \lambda\mathbf{E}[X]$ for its offered load and c_X^2 for the squared coefficient of variation of X. Applying the Pollaczek–Khinchine formula (52) with these quantities gives the mean queue length, denoted by $L_q^{\mathrm{P\text{–}K}}(X)$. These auxiliary loads are not the busy fractions of the corresponding full models. In particular, ϱ_{T^*} may exceed ϱ, as in Table 2. The numerical results for the auxiliary queues, presented in Tables 1 and 2, clearly show the isolated

consequences of the local catastrophes for the performance measures. The observed non-monotonicity in these (Pollaczek-Khinchine!) results is not only due to smaller loads (i.e. ϱ_{S^*} and ϱ_{T^*}) for larger squared coefficients of variation c_S^2: it can be shown that also the squared coefficients of variation of the adapted service times themselves (i.e. $c_{S^*}^2$ and $c_{T^*}^2$) are non-monotonic in c_S^2 (we skip the mathematical details). Comparing each auxiliary queue with the corresponding full model reveals the additional effect of the possible occurrence of complete catastrophes.

These comparisons separate the different mechanisms, but they do not give an additive decomposition of the full-model queue length because the mechanisms interact. In particular, S^* and T^* are not direct measures of the duration of active service in the full models. There, an active service attempt lasts $\min\{S, E(\gamma)\}$ because both local and complete catastrophes terminate the service attempt.

Observed non-monotonicity. We use $c_S^2 \in \{0.6, 0.8, 1, 2, 5, 10, 20, 50, 100, 1000\}$. In Tables 1 and 2, bold type marks the largest tabulated entry in each column, except that it marks the smallest entry for $\hat{\mathcal{D}}(\alpha\gamma)$ and $\mathbf{P}$(fully s.).

	Auxiliary $M/G/1$ comparison queue				Full-model performance						
c_S^2	$\mathbf{E}[S^*]$	ϱ_{S^*}	$c_{S^*}^2$	$L_q^{\mathrm{P-K}}(S^*)$	L_q	$\mathbf{E}[D]$	$\mathbf{E}[D \mid \mathrm{s}]$	p_{00}	$\hat{\mathcal{D}}(\alpha\gamma)$	$\hat{\mathcal{F}}_S(\gamma)$	$\mathbf{P}$(fully s.)
0.6	**0.9804**	**0.8823**	0.6053	5.31	2.6323	3.2127	2.8727	0.1980	0.9269	0.9519	0.8823
0.8	0.9780	0.8802	0.8030	5.83	2.7837	3.4235	3.0286	0.2054	0.9227	0.9522	0.8785
1	0.9756	0.8780	1.0000	6.32	2.9206	3.6178	3.1678	0.2124	0.9189	0.9524	0.8751
2	0.9640	0.8676	1.9755	8.46	3.4492	4.4070	3.6862	0.2426	0.9042	0.9535	0.8621
5	0.9313	0.8382	4.8073	12.60	4.2562	5.8083	4.3673	0.3070	0.8818	0.9563	0.8433
10	0.8829	0.7947	9.2284	15.73	**4.6761**	6.8489	**4.5318**	0.3766	**0.8701**	0.9602	**0.8355**
20	0.8043	0.7238	17.0655	**17.14**	4.6276	**7.3802**	4.1145	0.4607	0.8715	0.9656	0.8415
50	0.6535	0.5882	34.3298	14.84	3.6842	6.6047	2.8343	0.5769	0.8977	0.9738	0.8742
100	0.5274	0.4747	**50.0398**	10.95	2.6136	5.0412	1.8149	0.6513	0.9274	0.9791	0.9080
1000	0.2907	0.2616	38.3632	1.82	0.4506	0.9178	0.2916	**0.7613**	0.9875	**0.9866**	0.9742

Table 1: Performance measures for the model without repair times ($\varrho = 0.90$, $\gamma = 0.05$, $\alpha = 0.5$, $\mathbf{E}[S] = 1$).

	Auxiliary $M/G/1$ comparison queue				Full-model performance						
c_S^2	$\mathbf{E}[T^*]$	ϱ_{T^*}	$c_{T^*}^2$	$L_q^{\mathrm{P-K}}(T^*)$	L_q	$\mathbf{E}[D]$	$\mathbf{E}[D \mid \mathrm{s}]$	p_{00}	$\hat{\mathcal{D}}(\alpha\gamma)$	$\hat{\mathcal{F}}_S(\gamma)$	$\mathbf{P}$(fully s.)
0.6	**1.0784**	**0.9706**	1.1337	**34.13**	5.0969	6.7346	5.5483	0.1087	0.8584	0.9519	0.8172
0.8	1.0758	0.9682	1.3151	34.12	5.1870	6.8851	5.6366	0.1155	0.8559	0.9522	0.8150
1	1.0732	0.9659	1.4959	34.09	5.2683	7.0244	5.7143	0.1220	0.8537	0.9524	0.8130
2	1.0604	0.9544	2.3908	33.83	5.5786	7.5942	5.9864	0.1506	0.8450	0.9535	0.8057
5	1.0244	0.9220	4.9893	32.63	6.0110	8.5984	**6.2283**	0.2144	0.8330	0.9563	**0.7967**
10	0.9712	0.8741	9.0474	30.49	**6.1204**	9.2743	6.0135	0.2864	**0.8300**	0.9602	0.7970
20	0.8847	0.7962	16.2453	26.82	5.7712	**9.3990**	5.2254	0.3764	0.8397	0.9656	0.8108
50	0.7189	0.6470	32.1297	19.64	4.4767	8.0518	3.5403	0.5048	0.8756	0.9738	0.8527
100	0.5802	0.5222	**46.6534**	13.60	3.1922	6.0853	2.3087	0.5889	0.9113	0.9791	0.8923
1000	0.3198	0.2878	37.0591	2.21	0.6969	1.2720	0.5222	**0.7156**	0.9806	**0.9866**	0.9675

Table 2: Performance measures for the model with repair times ($\varrho = 0.90$, $\gamma = 0.05$, $\alpha = 0.5$, $\nu = 0.25$, $\mathbf{E}[S] = 1$).

The most interesting observation from Tables 1 and 2 is that the mean queue lengths and waiting-time measures are not monotone in c_S^2. Without repair times, L_q reaches its largest tabulated value

4.6761 at $c_S^2 = 10$, while $\mathbb{E}[D]$ reaches its largest value 7.3802 at $c_S^2 = 20$. With repair times, the corresponding maxima are 6.1204 at $c_S^2 = 10$ and 9.3990 at $c_S^2 = 20$. The mean waiting time of a successful customer turns over at $c_S^2 = 10$ without repairs and at $c_S^2 = 5$ with repairs. Thus, the different performance measures need not reach their maxima at the same value of c_S^2.

Successful and fully successful customers. Combining Eqs. (18) and (51) for the models without and with repair times, respectively, with Little's Law $L_q = \lambda \overline{W}_q$, we find for both models that

$$\hat{\mathcal{D}}(\alpha\gamma) = 1 - \frac{\alpha\gamma}{\lambda} L_q.$$

Since λ, α, and γ are kept fixed in the numerical examples, the fraction of customers who successfully reach the server and start their service is an affine decreasing function of L_q. Hence, this fraction is smallest precisely when the mean queue length is largest.

The fraction of fully successful customers, i.e. customers who reach the server and subsequently complete their service without interruption by a catastrophe, is equal to

$$\mathbb{P}(\text{fully s.}) = \hat{\mathcal{D}}(\alpha\gamma)\hat{\mathcal{F}}_S(\gamma) = \left(1 - \frac{\alpha\gamma}{\lambda} L_q\right) \hat{\mathcal{F}}_S(\gamma). \tag{54}$$

In Tables 1 and 2, the probability of reaching the server, $\hat{\mathcal{D}}(\alpha\gamma)$, initially decreases and then increases with c_S^2. By contrast, the probability of completing service given that service has started, $\hat{\mathcal{F}}_S(\gamma)$, increases throughout both tables.

In Table 2, the increase in $\hat{\mathcal{F}}_S(\gamma)$ between $c_S^2 = 5$ and $c_S^2 = 10$ more than compensates for the decrease in $\hat{\mathcal{D}}(\alpha\gamma)$, so the fully successful fraction already increases while L_q still increases. Its smallest tabulated value occurs at $c_S^2 = 5$, whereas the mean queue length is largest at $c_S^2 = 10$.

The same expression also clarifies the effect of repairs. For each fixed c_S^2, the service-time distribution and catastrophe rate are identical in both models, so $\hat{\mathcal{F}}_S(\gamma)$ is the same. Repairs can delay the start of service, during which waiting customers remain exposed to complete catastrophes. Once service starts, however, the probability of completing is unchanged. Thus, the smaller fully successful fractions in Table 2, compared with Table 1, are entirely due to the smaller probabilities of reaching the server.

Two opposing effects of catastrophes. The fractions of successful and fully successful customers are thus determined by L_q and $\hat{\mathcal{F}}_S(\gamma)$. It remains to interpret why L_q and the waiting times first increase and then decrease with c_S^2. Under the gamma-normalisation, as c_S^2 grows, μ_1 increases towards $4/\mathbb{E}[S]$, whereas μ_2 and b decrease towards zero. Hence, an increasing fraction of the service times ends after the relatively fast first phase, while a decreasing fraction enters an increasingly slow second phase. In a classical $M/G/1$ queue, the rare but very long service times have a large adverse effect on congestion. In our models, however, these long services are likely to be interrupted by a catastrophe. The catastrophes therefore have two opposing effects on L_q and the waiting times. The first is the effect of variability: up to $c_S^2 = 100$, the duration of a service attempt still becomes more variable as c_S^2 grows, which tends to increase L_q. The second is the effect of the load: the catastrophes shorten the mean duration of a service attempt, which reduces the load on the server and tends to decrease L_q. For the service attempts in the full models, Table 3 below shows how both the mean and the squared coefficient of variation change with c_S^2. In the classical $M/G/1$ queue only the first effect is present, because the load stays equal to ϱ. In the auxiliary queues, the two effects can be separated exactly, because for $X \in \{S^*, T^*\}$ the Pollaczek–Khinchine formula (52) is the product of a load factor and a variability factor:

$$L_q^{\text{P–K}}(X) = \underbrace{\frac{\varrho_X^2}{2\,(1-\varrho_X)}}_{\text{load}} \cdot \underbrace{\left(1 + c_X^2\right)}_{\text{variability}}. \tag{55}$$

Tables 1 and 2 show both effects: the offered loads ϱ_{S^*} and ϱ_{T^*} decrease, while $c^2_{S^*}$ and $c^2_{T^*}$ first increase and then decrease. For small and moderate values of c^2_S, the effect of variability seems to dominate, and L_q, the waiting times and $L_q^{\text{P–K}}(S^*)$ increase. For large values, the effect of the load takes over, and they decrease. In the auxiliary queue with repairs, however, ϱ_{T^*} starts close to one, so that the effect of the load dominates throughout and $L_q^{\text{P–K}}(T^*)$ only decreases.

Duration of a service attempt. We now return to the full models to examine how service interruptions, complete queue clearings, and conditioning on success help interpret the numerical results in Tables 1 and 2. In the full models, both effects can be measured on the service attempt itself. An active service attempt ends after $\min\{S, E(\gamma)\}$ because both local and complete catastrophes interrupt service. Its expected duration is

$$\mathbf{E}[\min\{S, E(\gamma)\}] = \int_0^\infty \mathrm{e}^{-\gamma t}\mathbf{P}(S > t)\mathrm{d}t = \frac{1 - \hat{\mathcal{F}}_S(\gamma)}{\gamma}. \tag{56}$$

Thus, the column $\hat{\mathcal{F}}_S(\gamma)$ in Tables 1 and 2 also determines the mean time for which a customer who starts service actually occupies the server, excluding a possible subsequent repair. Its squared coefficient of variation follows from the formula for $c^2_{S^*}$ in the Appendix with θ replaced by γ. Since a repair starts only after the attempt has ended, both quantities are the same with and without repair times. Table 3 shows them next to L_q.

	Service attempt $\min\{S, E(\gamma)\}$		L_q	
c^2_S	mean	squared c.v.	without repairs	with repairs
0.6	**0.961**	0.61	2.6323	5.0969
0.8	0.957	0.81	2.7837	5.1870
1	0.952	1.00	2.9206	5.2683
2	0.931	1.95	3.4492	5.5786
5	0.873	4.63	4.2562	6.0110
10	0.796	8.55	**4.6761**	**6.1204**
20	0.688	14.76	4.6276	5.7712
50	0.524	25.13	3.6842	4.4767
100	0.418	**30.19**	2.6136	3.1922
1000	0.269	12.45	0.4506	0.6969

Table 3: Mean and squared coefficient of variation of the duration $\min\{S, E(\gamma)\}$ of a service attempt in the full models, next to the mean queue lengths of Tables 1 and 2 ($\varrho = 0.90$, $\gamma = 0.05$, $\alpha = 0.5$, $\mathbf{E}[S] = 1$).

Although $\mathbf{E}[S] = 1$ is kept fixed, the mean duration of a service attempt decreases by 72%, from 0.96 at $c^2_S = 0.6$ to 0.27 at $c^2_S = 1000$. Up to $c^2_S = 10$, however, it decreases by only 17%, while its squared coefficient of variation increases from 0.61 to 8.55: the effect of variability seems to dominate, and L_q increases. Beyond $c^2_S = 10$, the mean duration decreases by a further two thirds, while the variability grows much less and eventually decreases: the effect of the load takes over, and L_q decreases. Finally, since (56) gives $\hat{\mathcal{F}}_S(\gamma) = 1 - \gamma\mathbf{E}[\min\{S, E(\gamma)\}]$, the shorter service attempts also explain why the probability of completing a service, the factor $\hat{\mathcal{F}}_S(\gamma)$ in (54), increases throughout both tables.

Summarising, for the gamma-normalisation and parameter values in Tables 1 and 2, mean queue lengths and waiting times first increase and subsequently decrease with c^2_S. The auxiliary queues clarify the competing effects of decreasing offered loads and changing service-time variability after interruption.

In the full models, each service attempt ends either with service completion or interruption by a catastrophe. Both the mean duration of such an attempt and the expected repair time per started service attempt decrease, although $\mathbf{E}[S]$ remains fixed. The fraction of successful customers follows the opposite pattern to L_q through their exact relation. The fraction of fully successful customers also depends on the probability of completing service after reaching the server, which increases throughout the tables.

5 Final remarks

In this paper we have studied two versions of an $M/C_2/1$ queueing model with Bernoulli catastrophes, a simplified model without repair times, and a more complicated model with repair times. For both models we have presented the steady-state analysis in full detail from which we have derived explicit expressions for the queue-length probabilities, the mean queue length, the LST of the virtual waiting time, the fraction of successful customers, and the mean waiting time of a successful customer. Also the mean overall waiting time has been calculated, which resulted in a nice alternative expression for the fraction of successful customers. In our extensive numerical results we have observed surprising non-monotonicity results which also have been partially explained. At this stage it is not clear to the authors whether these non-monotonicity results are only due to the chosen gamma normalisation of the Coxian-2 distributed service times, hence maybe not a characteristic of the models themselves. In a subsequent paper we will pursue this topic further, using other service-time distributions. One of the main messages of the paper is that the 'Pollaczek-Khinchine intuition' that more irregularity of the service times brings with it a deteriorating overall performance of $M/G/1$-like queueing models is not always true: enriching an $M/G/1$ model with extra features (e.g. vacations, catastrophes, impatience) sometimes contradicts this widely established intuition, as we have shown in this paper.

References

[1] Eitan Altman and Uri Yechiali (2006). Analysis of customers' impatience in queues with server vacations, *Queueing Systems: Theory and Applications* **52**(4): 261–279.

[2] Sundar Rajan Balasubramanian, Ganesan V, Rita Samikannu (2022). Steady-State Behaviour of $M^X/M/G_\eta/1$ Queueing System with Catastrophe, *International Journal of Difference Equations* **17**(2): 189–198.

[3] A. di Crescenzo, V. Giorno, A.G. Nobile, L.M. Ricciardi (2003). On the $M/M/1$ Queue with Catastrophes and Its Continuous Approximation, *Queueing Systems: Theory and Applications* **43**: 329–347.

[4] Mustafa Demircioglu, Herwig Bruneel and Sabine Wittevrongel (2021). Analysis of a Discrete-Time Queueing Model with Disasters, in special issue *Recent Research in Queuing Theory and Stochastic Models* of *Mathematics* **9**(24): 3283, editors Ivan Atencia, José Luis Galán-García.

[5] Spiros Dimou and Antonis Economou (2013). The Single Server Queue with Catastrophes and Geometric Reneging, *Methodology and Computing in Applied Probability* **15**: 595–621.

[6] Antonis Economou and Stella Kapodistria (2010). Synchronized abandonments in a single server unreliable queue, *European Journal of Operational research* **203**: 143–155.

[7] Erol Gelenbe, Peter Glynn and Karl Sigman (1991). Queues with Negative Arrivals, *Journal of Applied Probability* **28**: 245–250.

[8] Veena Goswami (2014). Analysis of Impatient Customers in Queues with Bernoulli Schedule Working Vacations and Vacation Interruption, *Journal of Stochastics*, Volume **2014**, Article ID 207285 (10 pages).

[9] P.G. Harrison and E. Pitel (1993). Sojourn times in Single-Server Queues with Negative Customers, *Journal of Applied Probability* **30**: 943–963.

[10] Gautam Jain and Karl Sigman (1996). A Pollaczek-Khintchine Formula for $M/G/1$ Queues with Disasters, *Journal of Applied Probability* **33**(4): 1191–1200.

[11] Bo Keun Kim, Doo Ho Lee (2014). The M/G/1 queue with disasters and working breakdowns, *Applied Mathematical Modelling* **38**: 1788-1798.

[12] B. Krishna Kumar, D. Arivudainambi (2000). Transient Solution of an $M/M/1$ Queue with Catastrophes, *Computers and Mathematics with Applications* **40**: 1233–1240.

[13] B. Krishna Kumar, A. Krishnamoorthy, S. Pavai Madheswari, S. Sadiq Basha (2007). Transient analysis of a single server queue with catastrophes, failures and repairs, *Queueing Systems: Theory and Applications* **56**: 133–141.

[14] B. Krishna Kumar, S.R. Anantha Lakshmi, S. Anbarasu, S. Pavai Madheswari (2014). Transient and steady state analysis of queueing systems with catastrophes and impatient customers, *Int. J. Mathematics in Operational Research* **6, No. 5**: 523–549.

[15] Young Liu, Liwei Liu (2023?). an $M/PH/1$ Queue with Catastrophes, preprint, *Research Square*.

[16] Shakir Majid (2023). Performance Analysis of a Markovian Queue with Impatient Customers and Working Vacation, *Journal of the Operations Society of China* **11**: 133–156.

[17] Ramachandran Remya, Jaisingh Ebenesar Anna Bagyam (2024). Comprehensive Discussion of the Repairable Single Server Catastrophe and Multiple Vacation Queueing Model, *Suranaree J. Sci. Technol.* **31(2)**: 1–8.

[18] K. Thilagavathy, G. Ayyapan (2025). Investigating the Performance of $MAP/PH/1$ Queueing Systems with Catastrophic Failures, Negative Arrivals, Server Breakdowns, Repairs and Lack of Customer Tolerance for Delays, *Reliability: Theory & Applications*, **20**, No 4 (89): 852–866.

[19] Naishuo Tian and Zhe George Zheng (2006). *Vacation Queueing Models: Theory and Applications*, Springer Science+Business Media, LLC.

[20] Henk C. Tijms (2003) *A First Course in Stochastic Models*, John Wiley & Sons Ltd.

[21] K.V. Vijayashree and K. Ambika (2020). An $M/M/1$ Queueing Model Subject to Differentiated Working Vacation and Customer Impatience, in S. Balusamy et al.(Eds.) *Computational Intelligence, Cyber Security and Computational Models. Models and Techniques for Intelligent Systems and Automation. ICC3 2019,* Springer Nature Singapore Pte Ltd.

[22] Uri Yechiali (2007). Queues with system disasters and impatient customers when sytem is down, *Queueing Systems: Theory and Applications* **56**: 195–202.