symmetry

MDPI

Article

Response Time of Queueing Mechanisms

Andrzej Chydzinski * and Blazej Adamczyk

Department of Computer Networks and Systems, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
* Correspondence: andrzej.chydzinski@polsl.pl

Abstract: We study response time, a key performance characteristic of queueing mechanisms. The studied model incorporates both active and passive queue management, arbitrary service time distribution, as well as a complex model of arrivals. Therefore, the obtained formulas can be used to calculate the response time of many real queueing mechanisms with different features, by parameterizing adequately the general model considered here. The paper consists of two parts. In the first, mathematical part, we derive the distribution function for the response time, its density, and the mean value. This is done by constructing two systems of integral equations, for the distribution function and the mean value, respectively, and solving these systems with transform techniques. All the characteristics are derived both in the time-dependent and steady-state cases. In the second part, we present numerical values of the response time for a few system parameterizations and point out several of its properties, some rather counterintuitive.

Keywords: queueing system; response time; time-dependent analysis; steady-state analysis

1. Introduction

By the response time of a device or system, we mean the overall time a task (customer, job, networking packet, etc.) occupies the system, from entering the system until its completion. The response time of a queuing system is, by this definition, the time spent by a task in the queue while waiting for service/execution, plus the service time of this task.

The response time of most queuing systems is random in nature. Therefore, it can be fully characterized by its probability distribution or roughly by its moments (the mean, variance, etc.).

The mean response time is the principal characteristic of the performance of a queuing system (see, e.g., [1]). If we have to describe roughly the performance of a system by giving only one number, the best candidate for this number is the mean response time.

Therefore, studies on the distribution and the mean response time are of great importance. A few such studies have been carried out to date; the specifics are to be seen in the subsequent section. In short, all are devoted to queuing systems possessing some important features and mechanisms, while lacking others.

In reality, queuing systems may possess many different features and mechanisms. For instance, the arrival process may be of the Poisson type or the renewal type with arbitrary interarrival distribution. It may or may not incorporate group arrivals. The interarrival times may or may not be correlated. There may or may not be some other subtle effects present, e.g., correlation between the temporary intensity and the size of arriving groups, or correlation between sizes of groups. The service time may be of some specific type or of a general type. The queue management may be of a simple type (tail-drop) or of a more advanced type (active management).

The goal of this article is to calculate the response time in one universal queuing model, incorporating all the aforementioned possible features together. Moreover, the obtained results cover both the steady-state and time-dependent response time, as well as its distribution function, probability density, and the mean value.

check for updates

Citation: Chydzinski, A.; Adamczyk, B. Response Time of Queueing Mechanisms. *Symmetry* **2024**, *16*, 271. https://doi.org/10.3390/sym16030271

Academic Editor: Theodore E. Simos

Received: 30 January 2024
Revised: 20 February 2024
Accepted: 22 February 2024
Published: 24 February 2024

Specifically, the time-dependent response time is the overall time a task arriving at a particular time t occupies the system. Therefore, it is a function of t. In practice, it is calculated for small values of t to characterize the behavior of the system shortly after its activation, when its operation is deeply influenced by its initial state. The steady-state response time characterizes the stable work of the system, i.e., for $t \to \infty$, when the influence of the initial condition vanishes.

Active queue management, built into the examined model, is the mechanism that does not permit some arriving tasks to queue up, even if the waiting room is not filled completely. In some active management mechanisms, the decision about queuing up a new task is deterministic, based on some more or less advanced calculations, see e.g., [2]. In most algorithms, however, the decision about each task is random, i.e., permission to queue up happens with some probability. This probability may change in time, being influenced by the system's state. Furthermore, it is often assumed that this probability is dependent on the number of tasks in the system. We incorporate such a mechanism in the analyzed model. It has several important applications listed below and is, in fact, an extension of passive queue management.

By passive management (or tail drop), we mean a natural permission policy, in which a new task can queue up if and only if there is free space in the waiting room. The policy investigated here can be reduced to passive management by using the task-rejection probability represented by the Heaviside step function. In general, however, this probability can be arbitrarily related to the number of tasks in the system—tasks can be rejected long before the waiting room becomes filled completely.

Active management considered here originated in computer networking, where it was proposed to organize queues of packets in packet switching devices [3–10]. Nonetheless, the applicability of the model is much wider than that. This is because exactly the same model portrays the natural inclination of people to resign from joining a queue with a probability dependent on the queue length. The queue could be, for example, a physical queue in an amusement park [11], a virtual queue in a call center [12], a jam of cars on a motorway, and many others.

The contribution of this article is mathematical and numerical. The mathematical contribution consists of 4 new theorems devoted to the response time, presenting in particular:

- the response time distribution function in the time-dependent case (Theorem 1),
- the response time probability density in the time-dependent case (Theorem 2),
- the mean response time in the time-dependent case (Theorem 3),
- the distribution function, density and mean response time in steady state (Theorem 4).

In the numerical part, we show mean response times and densities computed for several system parameterizations, both in the time-dependent and steady-state regimes. We pay attention to how these characteristics vary with load, with the initial conditions, with the presence or lack of autocorrelation, and with task-rejection probabilities. Some unexpected behavior of the response time is noticed.

The arrival process is represented here by the batch Markovian arrival process, whose description and parameterization in the currently used form were first given in [13]. It is of utmost importance that this process can mimic many properties and nuances of real-life point processes. For example, it can imitate the interarrival distribution with arbitrary accuracy, the interarrival autocorrelation function, the group structure, the correlation between temporary arrival rate and the group size, correlation between sizes of groups, and others. Due to these qualities, it has been used in models of operation of various systems, including computer networking and telecommunications (see [14] and references therein), vehicular traffic [15,16], inventory systems [17,18], supply and maintenance systems [19,20], disaster models [21], and others.

The examined queuing model is asymmetric in the sense that the interarrival times can be autocorrelated, while the service times cannot. Such asymmetry is a trade-off between the applicability of the model and its analytical complexity. This will be debated further in Section 3.

The mathematical approach of this paper is based on constructing systems of integral equations and solving them with transform techniques. Specifically, two such systems of equations are proposed and solved: for the time-dependent distribution function of the response time (Equations (8) and (12)) and for the time-dependent mean response time (Equations (45) and (48)). The remaining characteristics of interest, i.e., densities and steady-state response times, are derived from the solutions of these two systems of equations.

The rest of the article develops as follows. In Section 2, the literature on the response time and related queuing models is reviewed. In Section 3, the modeling framework is presented, including details of the arrival process and organization of the queuing system with its active management. Several important special cases are discussed as well. Section 4 houses the paper's key contribution, namely theorems devoted to the response time distribution function, its probability density and mean value, and steady state. In Section 5, numerical results are provided and commented on. Finally, closing remarks are given in Section 6, as well as suggestions for future work.

2. Related Work

As far as the authors know, the results of this article are new.

All the previous studies of the response time, or the related waiting time, lack one or more important features of the model examined herein, such as active management, a general type of service distribution, autocorrelated arrivals, or group arrivals. Moreover, in most papers, the analysis is reduced to the steady state only, with no time-dependent solution.

Specifically, derivations of the response time or waiting time of classic queuing models, such as M/M/1, M/G/1, and G/M/1 types, can be found in many queuing monographs, e.g., [1,22,23]. Unfortunately, these models include neither active management nor autocorrelation.

Perhaps the closest analysis is performed in [13,24], where the waiting time is computed for the model with the same arrival and service processes as herein. Specifically, the steady-state waiting time is studied in [13], while the time-dependent one in [13]. The analysis is then continued in the monograph [14], p. 161. However, the model of [13,14,24] does not incorporate a critical component examined here, i.e., active queue management. This feature significantly expands universality and applicability of the model but also requires a different analytical approach.

Conversely, there are several studies with active management in the queuing model, but without any derivations of the response time or the waiting time [25–30].

There are also other works with the same or similar model of the arrival process, but without active queue management [15–21,31–33].

Finally, there are papers that deal with the response or waiting time and do involve active queue management, but with a simplified arrival process, which either lacks group arrivals [34] or autocorrelation [35]. A general arrival process with powerful modeling capabilities is an essential feature of the model examined here.

3. Modeling Framework

We analyze the queuing model with a single service station. Namely, the arriving tasks form a queue in the waiting room in the order of arrival. Concurrently, the queue is served from the front by the service station. The service/execution time of a task is random and has a distribution function $F(x)$ of arbitrarily complex form. Service times of distinct tasks are mutually independent. The waiting room is finite and has a size of $K-1$. This means that the maximum number of tasks in the system can be K, encompassing the position for service. If upon a task arrival there are already K tasks in the system, the just-arrived task is not permitted to queue up. It exits the system unattended and never comes back.

Furthermore, any arriving task may be forbidden to queue up, even if the waiting room isn't filled completely. This happens with a probability of $d(n)$ for each arriving task,

where n denotes the number of tasks occupying the system on the arrival of this new task. This is the active management mechanism built into the model.

We see immediately that parameterizing function $d(n)$ we can obtain passive management of the system as well. Namely, if

$$d(n) = \begin{cases} 0, & \text{for} \quad n < K, \\ 1, & \text{for} \quad n \geq K, \end{cases}$$

then a task can queue up only if the waiting room is not filled completely. This is the passive management case. If, on the other side, $d(n) \in (0,1)$ for some $n < K$, then active management takes place. In the examined model, $d(n)$ can have any form, so the derived formulas are applicable to systems with passive and active management.

The queue receives arrivals according to the batch Markovian arrival process [13]. The evolution of this process is governed by the modulating process $J(t)$, which is an m-state Markov chain with continuous time and rate matrix D. The modulating process controls arrivals of tasks, perhaps in groups.

In practice, the batch Markovian arrival process is parameterized by a sequence of $m \times m$ matrices: D_0, D_1, D_2, ... For consistency, D_k must be non-negative for every $k \geq 1$, the rows of $D = \sum_{i=0}^{\infty} D_i$ must sum to 0, and it must hold that $D \neq D_0$. An expanded account of characteristics and properties of this process is available in [14]. Its most important characteristic, the arrival rate, is equal to:

$$\lambda = \pi \sum_{i=1}^{\infty} i D_i \mathbf{1}, \tag{1}$$

where

$$\mathbf{1} = [1, \ldots, 1]^T, \tag{2}$$

and π is the steady-state vector for matrix D, fulfilling the system:

$$\pi D = [0, \ldots, 0], \quad \pi \mathbf{1} = 1. \tag{3}$$

The mean service time is denoted by S, i.e.,:

$$S = \int_0^{\infty} x \mathrm{d}F(x) = \int_0^{\infty} (1 - F(x)) \mathrm{d}x. \tag{4}$$

Load of the system is therefore:

$$\rho = \lambda S, \tag{5}$$

where λ is given in (1).

$N(t)$ represents the total count of tasks in the system at time t, embracing the one under service, if applicable.

Note that the considered queuing model does not possess the symmetry typical of classic queuing models, such as M/M/1 or G/G/1. In the classic models, the service and arrival processes are symmetric in the sense that all service times are uncorrelated and all interarrival times are uncorrelated. The asymmetry of the model here originates from the fact that arrival times can be highly correlated, while the service times cannot. Specifically, the k-lag correlation of the batch Markovian arrival process is as follows:

$$\rho_c(k) = \pi D_0 \mathbf{1} \cdot \frac{\xi D_0^{-2} (D - D_0) \left((-D_0^{-1}(D - D_0))^{k-1} - \mathbf{1}\xi \right) D_0^{-2} (D - D_0) \mathbf{1}}{2\pi D_0^{-1} \mathbf{1} + 1}, \tag{6}$$

where ξ is the steady-state vector for matrix $-D_0^{-1}(D - D_0) - I$ whereas I stands for $m \times m$ unit matrix. It is well established that $\rho_c(k)$ can have high values even for large k and the shape of $\rho_c(k)$ can be very well fitted to empirical autocorrelation functions (see, e.g., [36]).

On the other side, service times are not correlated in the examined model by definition.

Such asymmetry is a trade-off between the applicability of the model and its analytical complexity. Specifically, the arrival process is very often correlated in reality (e.g., [14,36,37]). Therefore, the applicability of symmetric models with no autocorrelation of both service and interarrival times is limited. On the other hand, correlated service times, although not impossible, are much rarer in reality than correlated arrivals. Therefore, the asymmetric model with correlated arrivals and uncorrelated services seems to constitute a reasonable choice.

Finally, by a proper parameterization of matrices D_0, D_1, $D_2, \ldots$, we can obtain many simpler processes, which can also be of interest, if all the features of the full parameterization are not necessary. Specifically:

- if $m = 1$ and $D_0 = -\lambda$, $D_1 = \lambda$, then we acquire the Poisson process,
- if $m = 1$, $D_0 = -\lambda$ and $j \geq 1$ $D_k = p_j\lambda$, then we acquire the compound Poisson process,
- if $D_0 = T$ and $D_1 = -T\mathbf{1}\alpha$, then we acquire the renewal process (uncorrelated arrivals) with phase-type interarrival distribution with parameters (α, T), which can approximate any renewal process with arbitrary precision,
- if $D_0 = T$ and $D_k = -p_kT\mathbf{1}\alpha$, $k \geq 1$, then we acquire the renewal process with group arrivals,
- if $D_k = \mathbf{0}$ for $k \geq 2$, then we acquire the Markovian arrival process, which has no groups, but all the remaining features of the original process,
- if $D_0 = Q - \Lambda$ and $D_1 = \Lambda$, then we acquire the Markov-modulated Poisson process, the simplest autocorrelated process without group arrivals,
- if $D_j = p_jD_1$, $j \geq 1$, then we acquire a process with correlated interarrival times, but uncorrelated sizes of groups.

The results obtained herein are valid for arbitrary matrices D_0, D_1, $D_2, \ldots$, so they can be employed in all the listed cases, and many other.

In Table 1, important model parameters and characteristics of interest are listed.

Table 1. Symbol list.

K	system capacity
D_k	arrival process parameters
$N(t)$	queue size at t
m	number of modulating states
$J(t)$	modulating state at t
$d(n)$	active management function
λ	arrival rate
$F(t)$	service time distribution function
S	mean service time
ρ	system load
$\overline{R}_{ni}(t,x)$	distribution function of the response time at t
$\overline{r}_{ni}(t,x)$	probability density of the response time at t
$\overline{T}_{ni}(t)$	mean response time at t
$\overline{R}^{\infty}(x)$	distribution function of the response time in steady state
$\overline{r}^{\infty}(x)$	probability density of the response time in steady state
$\overline{T}^{\infty}$	mean response time in steady state

4. Response Time

Let $\tau_{ni}(t)$ denote the system response time at time t, given that $N(0) = n$, $J(0) = i$. In other words, $\tau_{ni}(t)$ is the time it would take a task to pass the system, assuming it arrived

at t and was queued up by active queue management. Note that we consider a potential, not actual arrival, at an arbitrary t.

Denote:

$$\overline{R}_{ni}(t,x) = \mathbb{P}\{\tau_{ni}(t) < x | N(0) = n, J(0) = i\}, \quad t > 0,\ x > 0. \tag{7}$$

Therefore, $\overline{R}_{ni}(t,x)$ is the response time distribution function for time t, assuming initial conditions $N(0) = n$, $J(0) = i$.

As we can see, $\overline{R}_{ni}(t,x)$ is a time-dependent characteristic, i.e., a function of t. Computing $\overline{R}ni(t,x)$ for a small t, say $t = 10$, we can characterize the response time of the system shortly after its activation, when its operation depends severely on its initial state.

We start the analysis with the case $N(0) > 0$. Given this, we have:

$$\begin{aligned}\overline{R}_{ni}(t,x) = &\sum_{j=1}^{m}\sum_{l=0}^{K-n}\int_0^t \overline{h}_{ij}(n,l,z)\overline{R}_{n+l-1,j}(t-z,x)\mathrm{d}F(z)\\ &+\sum_{j=1}^{m}\sum_{l=0}^{K-n}\overline{h}_{ij}(n,l,t)\int_t^{\infty}F^{(n+l)*}(x+t-z)\mathrm{d}F(z), \quad 1 \le n \le K,\end{aligned} \tag{8}$$

where $F^{(k)*}$ denotes the k-fold convolution of the distribution function F with itself, while $\overline{h}_{ij}(n,l,z)$ is probability that l tasks will be let to the waiting room by time z, and the state at time z will be j, assuming no service will be completed by z, and that it was $N(0) = n$, $J(0) = i$.

Equation (8) is created from the total probability rule applied with respect to z, the service completion moment. In the first component of (8), it is assumed $t \ge z$. Under such assumption, the new task number at z is $n + l - 1$, because l new tasks are queued up by z and 1 task departs at z. Furthermore, the new modulating state at z is j. Therefore, beginning from z, the conditional probability that $\tau_{ni}(t) < x$ is $\overline{R}_{n+l-1,j}(t-z,x)$.

In the second component of (8), it is assumed $t < z$. Under such assumption, the number of tasks in the system at t is $n + l$. Therefore, we can compute directly $\tau_{ni}(t)$. Namely, $\tau_{ni}(t)$ consists of the residual time of service of the task occupying the service position at t. This residual time equals $z - t$. Then it consists of $n + l - 1$ complete services of tasks present at t. Furthermore, we have to add one more service time, for the task arriving at t. In total, $\tau_{ni}(t)$ consists of $n + l$ complete service times, plus one partial, of duration $z - t$. This gives $\mathbb{P}\{\tau_{ni}(t) < x\} = F^{(n+l)*}(x + t - z)$.

The active management mechanism is present in Equation (8) in function $\overline{h}_{ij}(n,l,z)$, defined above. Naturally, $\overline{h}_{ij}(n,l,z)$ depends on the parameterization of the active management mechanism, through the function $d(n)$. It will be shown later how $\overline{h}_{ij}(n,l,z)$ can be computed and how it depends on $d(n)$ (see Formulas (61)–(67)).

The second integral in (8) can be transformed as follows:

$$\begin{aligned}\int_t^{\infty}F^{(n+l)*}(x+t-z)\mathrm{d}F(z) &= \int_t^{t+x}F^{(n+l)*}(x+t-z)\mathrm{d}F(z)\\ &= \int_0^{x}F^{(n+l)*}(x-z)\mathrm{d}_zF(z+t).\end{aligned} \tag{9}$$

The first transformation in (9) follows from relation: $F^{(n+l)*}(t) = 0$ for $t < 0$. The second transformation is just a replacement of variables. Taking (9) into account, (8) is equivalent to:

$$\overline{R}_{ni}(t,x) = \sum_{j=1}^{m}\sum_{l=0}^{K-n}\int_0^t \overline{h}_{ij}(n,l,z)\overline{R}_{n+l-1,j}(t-z,x)\mathrm{d}F(z) + \sum_{j=1}^{m}\sum_{l=0}^{K-n}\overline{h}_{ij}(n,l,t)g(n+l,t,x), \tag{10}$$

with

$$g(k,t,x) = \int_0^x F^{(k)*}(x-z)\mathrm{d}_z F(z+t). \tag{11}$$

Now we move to the case $N(0) = 0$. Given this case, we have:

$$\overline{R}_{0i}(t,x) = \sum_{j=1}^{m}\sum_{k=0}^{\infty} p(i,j,k)\sum_{l=0}^{K} q(0,k,l)\int_0^t \overline{R}_{lj}(t-z,x)\mu_i e^{-\mu_i z}\mathrm{d}z + F(x)e^{-\mu_i t}, \tag{12}$$

where $q(n,k,l)$ is probability that from k tasks arriving in a group, l taks are queued up by active queue management, assuming n taks occupied the system on this arrival of a group.

Equation (12) is created from the total probability rule applied with respect to time z, which is now an arrival or alteration of the modulating state, or both. It is known that in the batch Markovian arrival proces, group arrival of length k happens with probability $p(i,j,k)$, after a time period that has exponential distribution with mean $1/\mu_i$, where:

$$p(i,j,k) = [D_k]_{i,j}/\mu_i, \qquad k > 0. \tag{13}$$

$$p(i,j,0) = [D_0]_{i,j}/\mu_i, \qquad i \neq j, \tag{14}$$

$$\mu_i = -[D_0]_{i,i}. \tag{15}$$

In other words, $p(i,j,k)$ describes transitions of the group arrival process, i.e., the probability that the next arrival will be of size k, perhaps accompanied with an alteration of state from i to j. Specifically, if $k = 0$, then merely a change of state occurs, devoid of an arrival. If $k > 0$ and $i = j$, then a group arrival takes place, without a change of state. If $k > 0$ and $i \neq j$, then a group arrival occurs with a change of state. Finally, simultaneous absence of arrival and absence of change of state is impossible, i.e., $p(i,i,0) = 0$.

In the first component of (12), it is assumed $t \geq z$. Under such assumption, the task number at z is l, and the new modulating state at z is j. Therefore, beginning from z, the conditional probability that $\tau_{0i}(t) < x$ is $\overline{R}_{lj}(t-z,x)$. In the second component of (12) it is assumed $t < z$, which has probability $e^{-\mu_i t}$. Under assumption $t < z$, there are still 0 tasks in the system at t. Hence $\mathbb{P}\{\tau_{0i}(t) < x\} = F(x)$.

The active management mechanism is present in Equation (12) in function $q(0,k,l)$, defined above. It will be shown later how $q(n,k,l)$ can be derived (see Formulas (61)–(64)).

Note that (10) and (12) make up a system of $(K+1)m$ integral equations of Volterra type, for unknowns $\overline{R}_{ni}(t,x)$, where $0 \leq n \leq K$ and $1 \leq i \leq m$. We will solve this system now using the transform technique.

Denoting:

$$R_{ni}(s,x) = \int_0^\infty e^{-st}\overline{R}_{ni}(t,x)\mathrm{d}t, \tag{16}$$

from (10) we have:

$$R_{ni}(s,x) = \sum_{j=1}^{m}\sum_{l=0}^{K-n} U_{ij}(n,l,s)R_{n+l-1,j}(s,x) + \sum_{j=1}^{m}\sum_{l=0}^{K-n} W_{ij}(n,l,s,x), \quad 1 \leq n \leq K, \tag{17}$$

$$U_{ij}(n,l,s) = \int_0^\infty e^{-st}\overline{h}_{ij}(n,l,z)\mathrm{d}F(t), \tag{18}$$

$$W_{ij}(n,l,s,x) = \int_0^\infty e^{-st}\overline{h}_{ij}(n,l,t)g(n+l,t,x)\mathrm{d}t, \tag{19}$$

while from (12) we obtain:

$$R_{0i}(s,x) = \mu_i(\mu_i+s)^{-1}\sum_{j=1}^{m}\sum_{k=0}^{\infty}p(i,j,k)\sum_{l=0}^{K}q(0,k,l)R_{lj}(s,x) + F(x)(\mu_i+s)^{-1}, \tag{20}$$

Then, (17) yields:

$$R_n(s,x) = \sum_{l=0}^{K-n}U(n,l,s)R_{n+l-1}(s,x) + \sum_{l=0}^{K-n}W(n,l,s,x)\mathbf{1},\ \ 1 \le n \le K, \tag{21}$$

where

$$R_n(s,x) = [R_{n1}(s,x),\ldots,R_{nm}(s,x)]^T, \tag{22}$$

$$U(n,l,s) = \left[U_{ij}(n,l,s)\right]_{i,j=1,\ldots,m}, \tag{23}$$

$$W(n,l,s,x) = \left[W_{ij}(n,l,s,x)\right]_{i,j=1,\ldots,m}, \tag{24}$$

while (20) gives:

$$R_0(s,x) = \sum_{k=0}^{\infty}\sum_{l=0}^{K}q(0,k,l)M(k,s)R_l(s,x) + F(x)G(s), \tag{25}$$

with

$$M(k,s) = \left[\mu_i(\mu_i+s)^{-1}p(i,j,k)\right]_{i,j=1,\ldots,m}, \tag{26}$$

$$G(s) = \left[(\mu_1+s)^{-1},\ldots,(\mu_m+s)^{-1}\right]^T. \tag{27}$$

Lastly, denoting:

$$R(s,x) = [R_{01}(s,x),\ldots,R_{0m}(s,x),R_{11}(s,x),\ldots,R_{1m}(s,x),\ldots,R_{K1}(s,x),\ldots,R_{Km}(s,x)]^T, \tag{28}$$

we can obtain the solution of (21) and (25) using standard algebraic manipulations. The result is summarized in the subsequent theorem.

Theorem 1. *Transform of the distribution function of time-dependent response time equals:*

$$R(s,x) = P^{-1}(s)Q(s,x), \tag{29}$$

with

$$P(s) = [P_{ij}(s)]_{i,j=0,\ldots,K}, \tag{30}$$

$$P_{ij}(s,x) = \begin{cases} \sum_{k=0}^{\infty}q(0,k,j)M(k,s) - I, & \text{if } i=j=0,\\ \sum_{k=0}^{\infty}q(0,k,j)M(k,s), & \text{if } i=0, j>0\\ U(i,j-i+1,s) - I, & \text{if } 0<i<K, i=j,\\ U(i,j-i+1,s), & \text{if } 0<i\le j+1<K+1, i\ne j,\\ -I, & \text{if } i=j=K,\\ \mathbf{0}, & \text{otherwise,}\end{cases} \tag{31}$$

where $\mathbf{0}$ *is the* $m \times m$ *zero matrix and*

$$Q(s,x) = \left[-F(x)G(s)^T, V(1,s,x)^T,\ldots,V(K,s,x)^T\right]^T, \tag{32}$$

$$V(n,s,x) = -\sum_{l=0}^{K-n} W(n,l,s,x)\mathbf{1}, \tag{33}$$

while $M(k,s)$, $G(s)$, $U(n,l,s)$ *and* $W(n,l,s,x)$ *are defined in (26), (27), (18) and (19), respectively.*

If the service distribution is absolutely continuous and has density $f(x)$, then the response time at t is also absolutely continuous and has density $\bar{r}_{ni}(t,x)$:

$$\bar{r}_{ni}(t,x)\mathrm{d}x = \mathbb{P}\{\tau_{n,i}(t) \in \mathrm{d}x | N(0) = n, J(0) = i\}, \quad t > 0,\ x > 0. \tag{34}$$

Transform of this density is equal to:

$$r(s,x) = \frac{\partial R(s,x)}{\partial x}, \tag{35}$$

where

$$r(s,x) = [r_{01}(s,x), \ldots, r_{0m}(s,x), r_{11}(s,x), \ldots, r_{1m}(s,x), \ldots, r_{K1}(s,x), \ldots, r_{Km}(s,x)]^T, \tag{36}$$

$$r_{ni}(s,x) = \int_0^\infty e^{-st}\bar{r}_{ni}(t,x)\mathrm{d}t. \tag{37}$$

To find (35), we have to calculate the derivative of $Q(s,x)$, which requires the derivative of $V(n,s,x)$, which in turn requires the derivative of $W(n,l,s,x)$, which requires the derivative of $g(k,t,x)$. In the last two steps, we need to employ the Leibniz integral rule, necessary for differentiation under integrals. By calculating all these derivatives, we get the subsequent theorem.

Theorem 2. *If the service distribution has density* $f(x)$*, then the transform of density of the time-dependent response time equals:*

$$r(s,x) = P^{-1}(s)z(s,x), \tag{38}$$

where

$$z(s,x) = \left[-f(x)G(s)^T, v(1,s,x)^T, \ldots, v(K,s,x)^T\right]^T, \tag{39}$$

$$v(n,s,x) = -\sum_{l=0}^{K-n} w(n,l,s,x)\mathbf{1}, \tag{40}$$

$$w(n,l,s,x) = \left[\int_0^\infty e^{-st}\overline{h}_{ij}(n,l,t)y(n+l,t,x)dt\right]_{i,j=1,\ldots,m}, \tag{41}$$

$$y(k,t,x) = \int_0^x f^{(k)*}(x-z)f(z+t)dz. \tag{42}$$

For practical purposes, it suffices often to know the mean response time, instead of its full distribution. Let $\overline{T}_{ni}(t)$ denote the mean response time at t, given the initial state was $N(0) = n$, $J(0) = i$, i.e.,:

$$\overline{T}_{ni}(t) = \mathbb{E}\{\tau_{n,i}(t) | N(0) = n, J(0) = i\}, \quad t > 0,\ x > 0. \tag{43}$$

There are at least two ways, how the mean response time can be computed. We can either use Theorem 1 with the relation:

$$\overline{T}_{ni}(t) = \int_0^\infty \left(1 - \overline{R}_{ni}(t,x)\right)\mathrm{d}x, \tag{44}$$

or build equivalents of Equations (10) and (12) for $\overline{T}_{ni}(t)$ and solve them. In fact, the latter way is very easy, because it is straightforward to build and solve analogical equations.

Namely, for $N(0) > 0$ we have:

$$\begin{aligned}\overline{T}_{ni}(t) = &\sum_{j=1}^{m}\sum_{l=0}^{K-n}\int_0^t \overline{h}_{ij}(n,l,z)\overline{T}_{n+l-1,j}(t-z)\mathrm{d}F(z)\\ &+\sum_{j=1}^{m}\sum_{l=0}^{K-n}\overline{h}_{ij}(n,l,t)\int_t^{\infty}[(n+l)S+t-z)]\mathrm{d}F(z), \quad 1\le n\le K,\end{aligned} \tag{45}$$

which is built analogously as (10). Splitting and rearranging the second integral in (45) we get:

$$\begin{aligned}\overline{T}_{ni}(t) = &\sum_{j=1}^{m}\sum_{l=0}^{K-n}\int_0^t \overline{h}_{ij}(n,l,z)\overline{T}_{n+l-1,j}(t-z)\mathrm{d}F(z)\\ &+(1-F(t))S\sum_{j=1}^{m}\sum_{l=0}^{K-n}(n+l)\overline{h}_{ij}(n,l,t),\\ &+a(t)\sum_{j=1}^{m}\sum_{l=0}^{K-n}\overline{h}_{ij}(n,l,t), \quad 1\le n\le K,\end{aligned} \tag{46}$$

where

$$a(t) = \int_0^{\infty}[1-F(z+t)]\mathrm{d}z. \tag{47}$$

For $N(0) = 0$ we have:

$$\overline{T}_{0i}(t) = \sum_{j=1}^{m}\sum_{k=0}^{\infty}p(i,j,k)\sum_{l=0}^{K}q(0,k,l)\int_0^t \overline{T}_{lj}(t-z)\mu_i e^{-\mu_i z}\mathrm{d}z + Se^{-\mu_i t}, \tag{48}$$

which is built analogously as (12). Denoting:

$$T_{ni}(s) = \int_0^{\infty} e^{-st}\overline{T}_{ni}(t)\mathrm{d}t, \quad T_n(s) = [T_{n1}(s),\ldots,T_{nm}(s)]^T, \tag{49}$$

(46) yields:

$$T_n(s) = \sum_{l=0}^{K-n}U(n,l,s)T_{n+l-1}(s) + S\sum_{l=0}^{K-n}(n+l)Y(n,l,s)\mathbf{1} + \sum_{l=0}^{K-n}Z(n,l,s)\mathbf{1}, \quad 1\le n\le K, \tag{50}$$

where

$$Y(n,l,s) = \left[\int_0^{\infty} e^{-st}(1-F(t))\overline{h}_{ij}(n,l,t)\mathrm{d}t)\right]_{i,j=1,\ldots,m}, \tag{51}$$

$$Z(n,l,s) = \left[\int_0^{\infty} e^{-st}a(t)\overline{h}_{ij}(n,l,t)\mathrm{d}t)\right]_{i,j=1,\ldots,m}. \tag{52}$$

From (48) we get:

$$T_0(s) = \sum_{k=0}^{\infty}\sum_{l=0}^{K}q(0,k,l)M(k,s)T_l(s) + SG(s). \tag{53}$$

Using notation:

$$T(s) = [T_{01}(s),\ldots,T_{0m}(s),T_{11}(s),\ldots,T_{1m}(s),\ldots,T_{K1}(s),\ldots,T_{Km}(s)]^T, \tag{54}$$

and solving the system (50) and (53), we get the subsequent theorem.

Theorem 3. *Transform of the time-dependent mean response time equals:*

$$T(s) = P^{-1}(s)H(s), \tag{55}$$

where

$$H(s) = \left[-SG(s)^T, L(1,s)^T, \ldots, L(K,s)^T\right]^T, \tag{56}$$

$$L(n,s) = -S\sum_{l=0}^{K-n}(n+l)Y(n,l,s)\mathbf{1} - \sum_{l=0}^{K-n}Z(n,l,s)\mathbf{1}, \tag{57}$$

while $P(s)$, $G(s)$, $Y(n,l,s)$ and $Z(n,l,s)$ are defined in (31), (27), (51) and (52), respectively.

Now we may obtain the response time for the system in steady state ($t \to \infty$). Let $\overline{R}^{\infty}(x)$ be the response time distribution function, $\overline{r}^{\infty}(x)$ the response time density and $\overline{T}^{\infty}$ the mean response time, all of them for the system in steady state.

All of these characteristics can be calculated easily from Theorems 1–3. It is done by applying the final-value theorem (see p. 89 of [38]). As the result, we get the theorem below.

Theorem 4. *In steady state, the response time distribution function and its mean value are as follows:*

$$\overline{R}^{\infty}(x) = \lim_{s\to 0+} sR_{01}(s,x), \tag{58}$$

$$\overline{T}^{\infty} = \lim_{s\to 0+} sT_{01}(s), \tag{59}$$

where $R_{01}(s,x)$ and $T_{01}(s)$ are given in Theorems 1 and 3, respectively. Moreover, if the service distribution is absolutely continuous, then the steady-state density of the response time equals:

$$\overline{r}^{\infty}(x) = \lim_{s\to 0+} sr_{01}(s,x), \tag{60}$$

where $r_{01}(s,x)$ is given in Theorem 2.

Note that the choice $N(0) = n = 0$, $J(0) = i = 1$ is arbitrary and does not influence the steady-state results. In fact, any other $N(0)$ and $J(0)$ can be chosen without changing the output.

Lastly, observe that to use Theorems 1–4 effectively, we must find functions $q(n,k,l)$ and $\overline{h}_{ij}(n,l,t)$. This can be realized using the method proposed in [28], which gives:

$$q(n,0,0) = 1, \qquad q(n,k,0) = d^k(n), \quad k > 0, \tag{61}$$

$$q(n,0,l) = 0, \quad l > 0, \tag{62}$$

$$q(n,k,l) = 0, \quad \min\{K-n,k\} < l, \tag{63}$$

$$q(n,k,l) = d(n)q(n,k-1,l) + (1-d(n))q(n+1,k-1,l-1), \quad k > 0,\ l > 0. \tag{64}$$

Denoting

$$h(n,l,s) = \left[h_{ij}(n,l,s)\right]_{i,j=1,\ldots,m}, \qquad h_{ij}(n,l,s) = \int_0^{\infty} e^{-st}\overline{h}_{ij}(n,l,t)\mathrm{d}t, \tag{65}$$

we obtain:

$$h(n,0,s) = (I - A(n,0,s))^{-1}\mathrm{diag}(G(s)), \tag{66}$$

$$h(n,l,s) = (I - A(n,0,s))^{-1}\sum_{j=1}^{l}A(n,j,s)h(n+j,l-j,s), \quad l > 0, \tag{67}$$

where

$$A(n,l,s) = \sum_{k=0}^{\infty} q(n,k,l)M(k,s). \tag{68}$$

5. Examples

The parameterization of task arrival process, utilized in this section, is as follows:

$$D_0 = \begin{bmatrix} -0.275697 & 0.262572 & 0.001878 \\ 0.007508 & -0.181924 & 0.003755 \\ 0.000483 & 0.000403 & -0.650183 \end{bmatrix}, \tag{69}$$

$$D_1 = \begin{bmatrix} 0.000468 & 0.001407 & 0.000937 \\ 0.041777 & 0.000468 & 0.000422 \\ 0.001622 & 0.000080 & 0.160628 \end{bmatrix}, \tag{70}$$

$$D_2 = \begin{bmatrix} 0.000625 & 0.001875 & 0.001250 \\ 0.055698 & 0.000625 & 0.000562 \\ 0.002162 & 0.000107 & 0.214155 \end{bmatrix}, \tag{71}$$

$$D_5 = \begin{bmatrix} 0.000780 & 0.002343 & 0.001562 \\ 0.069627 & 0.000780 & 0.000702 \\ 0.002703 & 0.000133 & 0.267707 \end{bmatrix}. \tag{72}$$

As can be noticed, the arrival process includes groups, with a maximum group size of 5. The mean group size is 3, and the arrival rate of groups is 0.333333. Thus, the overall arrival rate is 1. Moreover, this process is highly autocorrelated (see Figure 1). Its 1-lag autocorrelation is above 0.4 and remains significant up to about 100-lag.

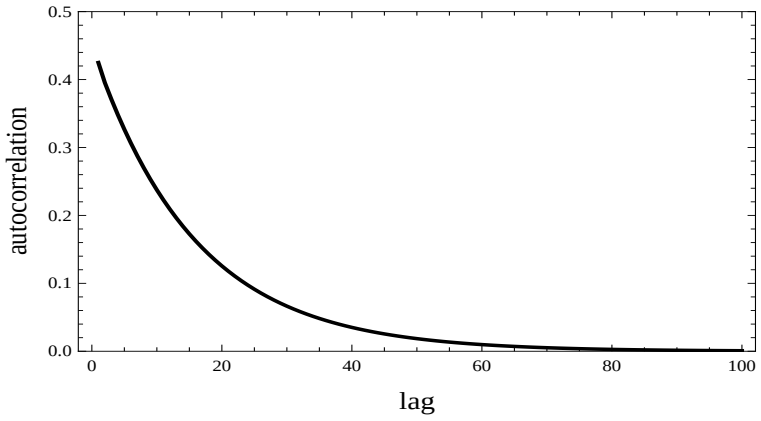

Figure 1. Correlation of interarrival times computed for arrival process (69)–(72).

The service distribution has the following density:

$$f(x) = \frac{1}{8\rho} e^{-\frac{1}{2\rho}x} + \frac{9}{8\rho} e^{-\frac{3}{2\rho}x}, \quad x > 0, \ \ \rho > 0, \tag{73}$$

where ρ is a parameter. For such density, the mean service time is:

$$S = \rho. \tag{74}$$

Hence, for $\lambda = 1$, the parameter ρ in (73) is equal to the load defined in (5). Distribution (73) has a rather moderate standard deviation, but greater than an exponential distribution would have. For example, for $\rho = 1$ it equals 1.29.

If not stated differently, the following function is used for letting tasks into the waiting room:

$$d(n) = \begin{cases} 0, & \text{for} \quad n < 15, \\ 0.0025(n-13)^3, & \text{for} \quad 15 \leq n < 20, \\ 1, & \text{for} \quad n \geq 20. \end{cases} \tag{75}$$

In Table 2, we have the mean response time computed for different moments in time and for steady state. Furthermore, three different initial numbers of tasks in the system ($N(0) = 0, 10, 20$) were assumed, with two values of load: full load, $\rho = 1$, and overload, $\rho = 1.5$. (Cases with $\rho < 1$ are less interesting, because the response time is low there for obvious reasons).

Table 2. The mean response time, $\overline{T}_{n1}(t)$, for various values of load, ρ, and initial numbers of tasks, $N(0)$.

	$\rho = 1$, $N(0) = 0$	$\rho = 1$, $N(0) = 10$	$\rho = 1$, $N(0) = 20$	$\rho = 1.5$, $N(0) = 0$	$\rho = 1.5$, $N(0) = 10$	$\rho = 1.5$, $N(0) = 20$
$t = 1$	1.50	10.1	20.0	2.45	15.6	30.5
$t = 2.5$	1.47	8.89	18.6	2.41	14.6	29.1
$t = 5$	1.75	7.17	16.6	2.91	13.2	27.2
$t = 10$	2.26	4.91	13.1	4.00	11.1	24.0
$t = 25$	3.40	3.77	6.20	6.42	8.82	16.9
$t = 50$	4.92	4.94	5.09	9.22	9.71	11.9
$t = 100$	6.85	6.86	6.87	12.6	12.6	12.7
$t = 250$	8.63	8.66	8.62	15.6	15.6	15.6
$t = 500$	8.92	8.82	8.94	16.1	16.1	16.1
$t = \infty$	8.94	8.94	8.94	16.1	16.1	16.1

In Figures 2–5, the response time density is depicted at different times, for 2 different values of ρ and 2 values of $N(0)$. Namely, in Figures 2 and 3, the fully loaded system is reflected, $\rho = 1$, with $N(0) = 10$ and $N(0) = 20$, respectively. In Figures 4 and 5, an overloaded queue is depicted, $\rho = 1.5$, again with $N(0) = 10$ and $N(0) = 20$, respectively.

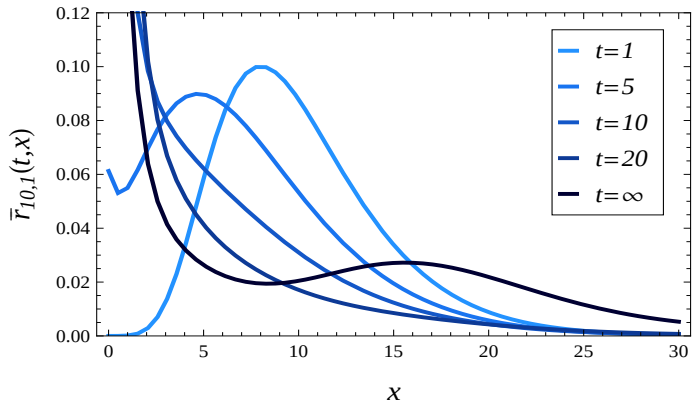

Figure 2. The response time density at different times assuming $N(0) = 10$ and $\rho = 1$.

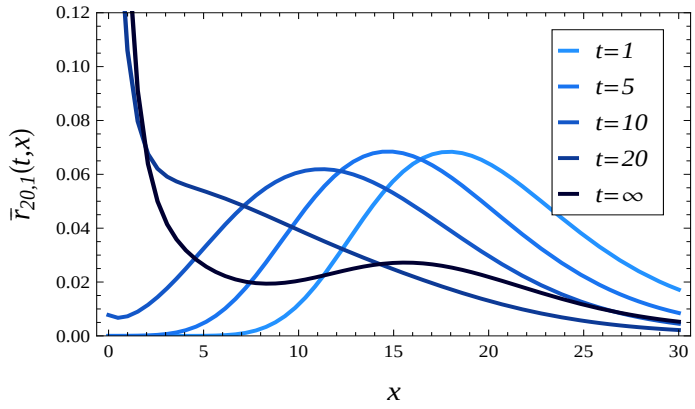

Figure 3. The response time density at different times assuming $N(0) = 20$ and $\rho = 1$.

In each of these figures, we may see the evolution of the density function towards the steady-state density (the darkest curve on each figure).

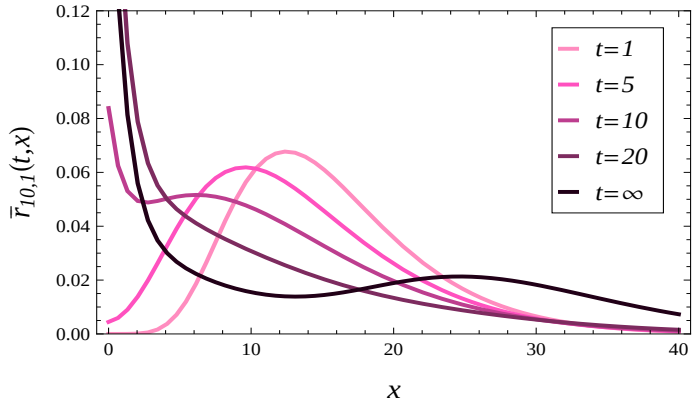

Figure 4. The response time density at different times assuming $N(0) = 10$ and $\rho = 1.5$.

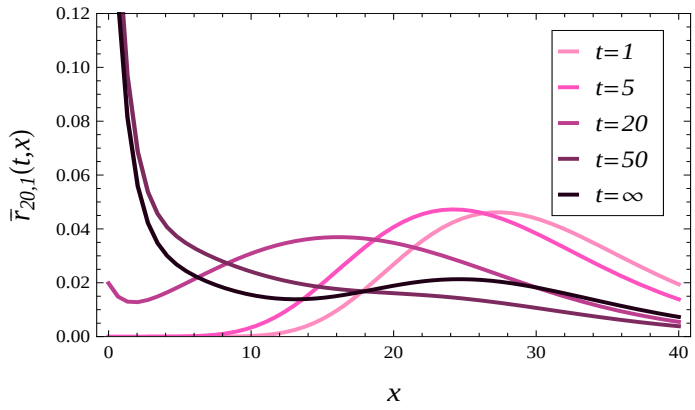

Figure 5. The response time density at different times assuming $N(0) = 20$ and $\rho = 1.5$.

5.1. Discussion of Results

As evident in Table 2, the mean response time approaches steady state when t reaches 500. (At $t = 250$, it is not stable yet, with a discrepancy of a few percent from the steady-state value.) 500 is a pretty long convergence time, most likely caused by the autocorrelation of arrivals.

We can also observe in Table 2 that the mean response time is often non-monotonic in time. Check, for instance, the case $N(0) = 10$, $\rho = 1$, where the response time starts at 10.1, then drops to 3.77, then grows to 8.94. Similarly in other cases.

What is surprising in Figures 2–5 is the high concentration of probability mass near $x = 0$ in steady state. It is especially unexpected in Figures 4 and 5, obtained for $\rho = 1.5$. One can expect that under such a high load, the queue should be long most of the time, so the response time should also be long and rarely close to zero. However, as we can see in Figures 4 and 5, this is not the case.

Such behavior of the response time is to be linked with the high autocorrelation of arrivals. To demonstrate that, we can compare response times obtained for autocorrelated arrivals (69)–(72) with results for uncorrelated arrivals, parameterized simply by: $D_0 = -1$, $D_1 = 1$.

5.2. Impact of Autocorrelation

In Figures 6 and 7, the density of the response time is shown at different times in overloaded queue with $\rho = 1.5$, and for $N(0) = 10$ and $N(0) = 20$, respectively. Both these figures were obtained for uncorrelated arrivals, $D_0 = -1$, $D_1 = 1$.

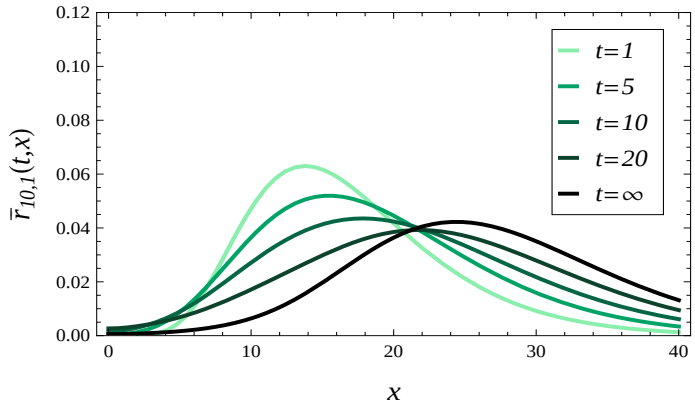

Figure 6. The response time density at different times assuming $N(0) = 10$, $\rho = 1.5$ and uncorrelated arrivals.

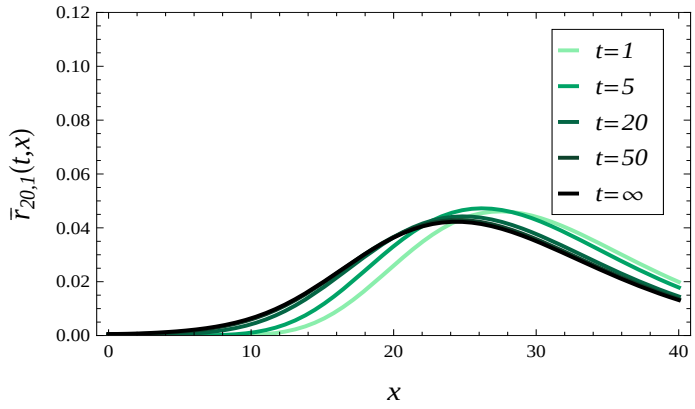

Figure 7. The response time density at different times assuming $N(0) = 20$, $\rho = 1.5$ and uncorrelated arrivals.

5.3. Discussion of Results

Figures 6 and 7 are to be compared with Figures 4 and 5, respectively.

In these pairs of figures, we may notice a very different behavior of the response time. Specifically, in Figures 4 and 5, the probability mass tends to concentrate around $x = 0$ as time passes. Conversely, in Figures 6 and 7, the probability mass concentrates around $x = 25$ as time goes on.

The effect of autocorrelation on the response time may be seen further in Figure 8. In this figure, the steady-state mean response time is shown depending on load, which

varies from 0 to 2. Specifically, the blue curve reflects the mean response time for autocorrelated arrivals (69)–(72), while the green curve reflects the response time for uncorrelated ones, $D_0 = -1$, $D_1 = 1$.

As it is clear in Figure 8, for a load below 0.95, the mean response time in the autocorrelated case is greater than in the uncorrelated one. Above 0.95, however, the situation reverses – the response time is greater if there is no correlation.

This surprising effect, visible also in Figures 4–7, can be interpreted as follows. Assume a high load, e.g., 1.5. In the correlated case, the arrival process is very irregular, i.e., it has periods of very low and very high arrival intensities. During a high-intensity phase, the queue length grows very quickly, due to the slow service. Consequently, active management kicks in aggressively, rejecting a large fraction of arriving tasks. In a low-intensity phase, a very short queue can be maintained, despite the slow service, because the arrival rate may be far less than 1 in such a phase.

In contrast, the arrival process without autocorrelation has a much more regular intensity. Therefore, the queue is moderate to long most of the time, and active management does not work so aggressively, as in the previous case.

Summarizing, when the load is high, in the autocorrelated case we have periods with short queues and periods with long queues (and extreme rejections). In the uncorrelated case, we have rather long queues most of the time (and moderate rejections). These factors make the response time shorter on average in the scenario with autocorrelated arrivals.

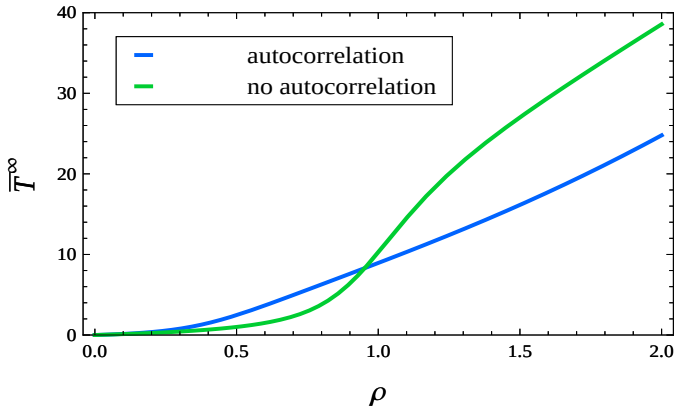

Figure 8. The mean response time in steady state versus load, for autocorrelated and uncorrelated arrivals.

5.4. *Impact of Active Management*

In the next group of numerical results, we will investigate the effect of parameterization of active management on the response time. In all the previous examples, we used $d(n)$ given in (75). Now the following formula will be utilized:

$$d_p(n) = d(n + p), \tag{76}$$

i.e., (75) shifted by a parameter p. Obviously, the greater p, the more sensitive active management, and the sooner it starts rejecting tasks.

In Figures 9 and 10, we can see the density of the response time in steady state for 5 values of parameter p. Figure 9 covers the case $\rho = 1$, while Figure 10 covers the case $\rho = 1.5$. In both figures, autocorrelated arrivals (69)–(72) are used.

As can be noticed in Figures 9 and 10, the greater p, the more probability mass is concentrated on low values of x, and the thinner the distribution tail.

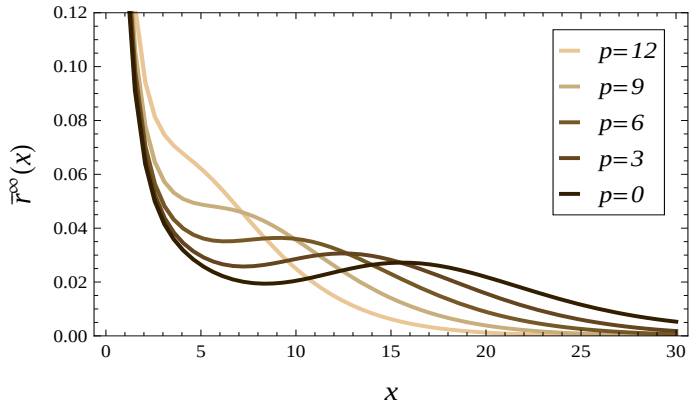

Figure 9. The steady-state response time density, for 5 values of parameter p in function $d_p(n)$ and $\rho = 1$.

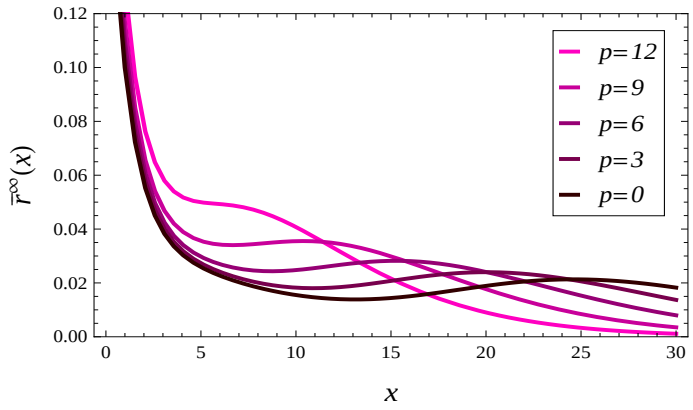

Figure 10. The steady-state response time density, for 5 values of parameters p in function $d_p(n)$ and $\rho = 1.5$.

5.5. *Impact of Service Time Distribution*

In the final set of numerical results, we will check the influence of the type of the service time distribution on the response time. So far, the hyperexponential service time (73) has been used in all numerical results. Now we will consider three more service time densities, namely:

- Pareto:

$$f_2(x) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}, \quad x > \beta, \tag{77}$$

  with parameters $\alpha = 3.236$, $\beta = 0.691$,
- Weibull:

$$f_3(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-(x/\beta)^{\alpha}}, \quad x > 0, \tag{78}$$

  with parameters $\alpha = 0.463$, $\beta = 0.430$,
- gamma:

$$f_4(x) = \beta^{\alpha} e^{-\beta x} x^{\alpha-1}/\Gamma(\alpha), \quad x > 0, \tag{79}$$

  with parameters $\alpha = 0.04$, $\beta = 0.04$.

All the distributions presented above have the mean value $S = 1.0$. They, however, differ by the standard deviation, which is 0.5, 2.5, and 5 for densities (77)–(79), respectively.

Furthermore, we will use not only bare distributions given in (77)–(79), but also their scaled variants $\rho\xi$, where $\rho > 0$ is a parameter, and ξ is a random variate with one of the distributions (77)–(79). Obviously, we have $S = \rho$ in every such case, so the factor ρ is equal to the load defined in (5).

5.6. Discussion of Results

In Figure 11, the mean response time in steady state is depicted for all four considered service time distributions as a function of ρ.

We can clearly observe two things. Firstly, the general shape of the curve seems not to be affected by the particular distribution used. The shape is roughly the same for all distributions (73), (77)–(79), even though densities (73), (77)–(79) differ significantly from each other.

Secondly, the slope of each curve is deeply affected by the standard deviation of the service time. The larger the standard deviation, the steeper the curve, regardless of the particular type of distribution. For $\rho = 1$, the considered standard deviations are 0.5, 1.29, 2.5 and 5, which is reflected in the same order of curves in Figure 11, from the least steep one (std. dev. of 0.5), to the most steep one (std. dev. of 5).

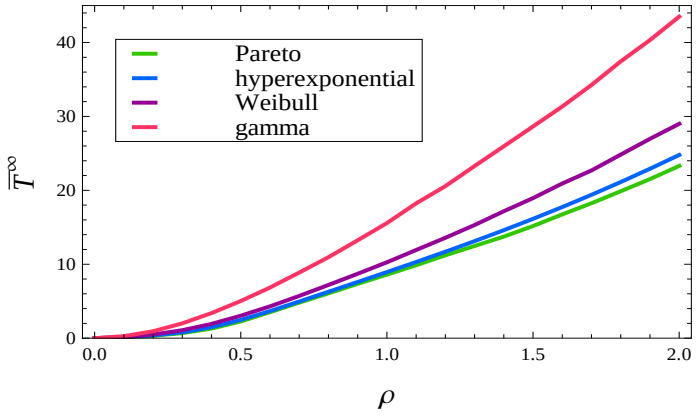

Figure 11. The mean response time in steady state versus load, for Pareto, hyperexponential, Weibull and gamma service time distribution.

6. Conclusions

We studied the response time using a queuing model which incorporates both active and passive queue management, arbitrary service distribution, and a complex model of arrivals, capable of mimicking arbitrary interarrival times, group arrivals, autocorrelated arrivals, correlated group sizes, and other advanced features.

Formulas for the distribution function, probability density, and the mean value were obtained both in the time-dependent case and steady state. These formulas can be used to characterize the response time of many queuing mechanisms, possessing any subset of the aforementioned features.

In numerical examples, we observed how the response time density and its mean value are influenced by load, autocorrelation, and active queue management parameterization, both in steady state and in the time-dependent scenarios.

The most unexpected observation made was that under high load, the mean response time can be significantly smaller when arrivals are autocorrelated than when they are uncorrelated. This contradicts a simplistic view that high autocorrelation of arrivals makes all the queuing characteristics worse. In fact, autocorrelation may improve the response time, perhaps at the cost of a decline in other characteristics.

The examined model, although rather general and broadly applicable, also has some limitations. Some of these limitations are consequences of the asymmetry between the arrival and service processes in the model.

The first limitation is that arrival times can be autocorrelated, but service times cannot. In reality, arrival processes are more often correlated than service processes, but correlated services can also be encountered.

The second limitation is that the arrival process possesses a group structure, whereas the service process is singular. Again, a singular type of service is common in many systems, but examples of systems with group service can be found as well.

The third limitation is the particular active management mechanism assumed in the study. In this mechanism, a task is forbidden to queue up based on the current system occupancy. This type of active management, although popular and intuitive, is not the only one possible. The results obtained here cannot be used for other active management types.

All these limitations may directly translate into propositions for future work. Firstly, the response time of a system with autocorrelated service times can be studied. Secondly, a model allowing simultaneous service of a group of arrivals can be investigated. Finally, a model with another type of active management can be studied mathematically, e.g., of the type proposed in [39].

Author Contributions: Conceptualization, A.C.; methodology, A.C. and B.A.; validation, A.C. and B.A.; formal analysis, A.C. and B.A.; investigation, A.C. and B.A.; writing—original draft preparation, A.C.; writing—review and editing, A.C. and B.A.; funding acquisition, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Centre, Poland, grant number 2020/39/B/ST6/00224.

Data Availability Statement: Data is contained within the article.

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

References

1. Kleinrock, L. *Queueing Systems: Theory*; John Wiley and Sons: New York, NY, USA, 1975.
2. Chrost, L.; Chydzinski, A. On the deterministic approach to active queue management. *Telecommun. Syst.* **2016**, *63*, 27–44. [CrossRef]
3. Floyd, S.; Jacobson, V. Random early detection gateways for congestion avoidance. *IEEE/Acm Trans. Netw.* **1993**, *1*, 397–413. [CrossRef]
4. Athuraliya, S.; Li, V.H.; Low, S.H.; Yin, Q. REM: Active queue management. *IEEE Netw.* **2001**, *15*, 48–53. [CrossRef]
5. Zhou, K.; Yeung, K.L.; Li, V. Nonlinear RED: Asimple yet efficient active queue management scheme. *Comput. Netw.* **2006**, *50*, 3784. [CrossRef]
6. Augustyn, D.R.; Domanski, A.; Domanska, J. A choice of optimal packet dropping function for active queue management. *Commun. Comput. Inf. Sci.* **2010**, *79*, 199–206.
7. Domanska, J.; Augustyn, D.; Domanski, A. The choice of optimal 3-rd order polynomial packet dropping function for NLRED in the presence of self-similar traffic. *Bull. Pol. Acad. Sci. Tech. Sci.* **2012**, *60*, 779–786. [CrossRef]
8. Giménez, A.; Murcia, M.A.; Amigó, J.M.; Martínez-Bonastre, O.; Valero, J. New RED-Type TCP-AQM Algorithms Based on Beta Distribution Drop Functions. *Appl. Sci.* **2022**, *12*, 11176. [CrossRef]
9. Feng, C.; Huang, L.; Xu, C.; Chang, Y. Congestion Control Scheme Performance Analysis Based on Nonlinear RED. *IEEE Syst. J.* **2017**, *11*, 2247–2254. [CrossRef]
10. Patel, S.; Karmeshu. A New Modified Dropping Function for Congested AQM Networks. *Wirel. Pers. Commun.* **2019**, *104*, 37–55. [CrossRef]
11. Doldo, P.; Pender, J.; Rand, R. Breaking the Symmetry in Queues with Delayed Information. *Int. J. Bifurc. Chaos* **2021**, *31*, 2130027. [CrossRef]
12. Rouba, I.; Mor, A.; Achal, B. Does the Past Predict the Future? The Case of Delay Announcements in Service Systems. *Manag. Sci.* **2016**, *63*, 6.
13. Lucantoni, D.M. New results on the single server queue with a batch Markovian arrival process. *Commun. Stat. Stoch. Model.* **1991**, *7*, 1–46. [CrossRef]
14. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer: Cham, Switzerland, 2020.
15. Alfa, A.S. Modelling traffic queues at a signalized intersection with vehicle-actuated control and Markovian arrival processes. *Comput. Math. Appl.* **1995**, *30*, 105. [CrossRef]
16. Alfa, A.S.; Neuts, M.F. Modelling vehicular traffic using the discrete time Markovian arrival process. *Transp. Sci.* **1995**, *29*, 109–117. [CrossRef]
17. Krishnamoorthy, A.; Joshua, A.N.; Kozyrev, D. Analysis of a Batch Arrival, Batch Service Queuing-Inventory System with Processing of Inventory While on Vacation. *Mathematics* **2021**, *9*, 419. [CrossRef]
18. Dudin, A.; Klimenok, V. Analysis of MAP/G/1 queue with inventory as the model of the node of wireless sensor network with energy harvesting. *Ann. Oper. Res.* **2023**, *331*, 839–866. [CrossRef]
19. Baek, J.W.; Lee, H.W.; Lee, S.W.; Ahn, S. A MAP-modulated fluid flow model with multiple vacations. *Ann. Oper. Res.* **2013**, *202*, 19–34. [CrossRef]

20. 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]
21. Dudin, A.; Karolik, A. BMAP/SM/1 queue with Markovian input of disasters and non-instantaneous recovery. *Perform. Eval.* **2001**, *45*, 19–32. [CrossRef]
22. Cohen, J.W. *The Single Server Queue*, Revised ed.; North-Holland Publishing Company: Amsterdam, The Netherlands, 1982.
23. Takagi, H. *Queueing Analysis—Finite Systems*; North-Holland: Amsterdam, The Netherlands, 1993.
24. Lucantoni, D.M.; Choudhury, G.L.; Whitt, W.The transient BMAP/G/1 queue. *Commun. Stat. Stoch. Model.* **1994**, *10*, 145–182. [CrossRef]
25. Hao, W.; Wei, Y. An Extended $GI^X/M/1/N$ Queueing Model for Evaluating the Performance of AQM Algorithms with Aggregate Traffic. *Lect. Notes Comput. Sci.* **2005**, *3619*, 395–414.
26. Kempa, W.M. Time-dependent queue-size distribution in the finite GI/M/1 model with AQM-type dropping. *Acta Electrotech. Inform.* **2013**, *13*, 85–90. [CrossRef]
27. Kempa, W.M. A direct approach to transient queue-size distribution in a finite-buffer queue with AQM. *Appl. Math. Inf. Sci.* **2013**, *7*, 909–915. [CrossRef]
28. Chydzinski, A.; Mrozowski, P. Queues with dropping functions and general arrival processes. *PLoS ONE* **2016**, *11*, e0150702. [CrossRef]
29. Tikhonenko, O.; Kempa, W.M. Performance evaluation of an M/G/N-type queue with bounded capacity and packet dropping. *Appl. Math. Comput. Sci.* **2016**, *26*, 841–854. [CrossRef]
30. Tikhonenko, O.; Kempa, W.M. Erlang service system with limited memory space under control of AQM mechanizm. *Commun. Comput. Inf. Sci.* **2017**, *718*, 366–379.
31. Chydzinski, A.; Adamczyk, B. Transient and stationary losses in a finite-buffer queue with batch arrivals. *Math. Probl. Eng.* **2012**, *2012*, 326830. [CrossRef]
32. Banik, A.D.; Chaudhry, M.L.; Wittevrongel, S.; Bruneel, H. A simple and efficient computing procedure of the stationary system-length distributions for GI^X/D/c and BMAP/D/c queues. *Comput. Oper. Res.* **2022**, *138*, 105564. [CrossRef]
33. Vishnevskii, V.M.; Dudin, A.N. Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks. *Autom. Remote Control* **2017**, *78*, 1361–1403. [CrossRef]
34. Chydzinski, A. Waiting Time in a General Active Queue Management Scheme. *IEEE Access* **2023**, *11*, 66535–66543. [CrossRef]
35. Chydzinski, A.; Adamczyk, B. Response time of the queue with the dropping function. *Appl. Math. Comput.* **2020**, *377*, 125164. [CrossRef]
36. Salvador, P.; Pacheco, A.; Valadas, R. Modeling IP traffic: Joint characterization of packet arrivals and packet sizes using BMAPs. *Comput. Netw.* **2004**, *44*, 335–352. [CrossRef]
37. Lel, W.; Taqqu, M.; Willinger, W.; Wilson, D. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Netw.* **1994**, *2*, 1–15.
38. Schiff, J.L. *The Laplace Transform: Theory and Applications*; Springer: New York, NY, USA, 1999.
39. Nichols, K.; Jacobson, V. Controlling Queue Delay. *Queue* **2012**, *55*, 42–50.