Stability analysis of multiserver discrete-time queueing systems with renewal-type server interruptions

Evsey Morozov[a,1], Dieter Fiems[*,b,2], Herwig Bruneel[b]

[a] *Institute of Applied Mathematical Research, Karelian Research Center, Russia*
[b] *SMACS Research Group, Department TELIN, Ghent University, Belgium*

Abstract

For many queueing systems the server is not continuously available. Service interruptions may result from repair times after server failures, from planned maintenance periods or from periods during which customers from other queues are being served. These service interruptions cause an overall performance degradation which is most striking when interruptions can start while a customer is being served and his service has to start all over after the interruption. This is the so-called preemptive repeat service discipline. This paper investigates stability conditions for discrete-time queueing systems with preemptive server interruptions. Under renewal assumptions for arrival, service and interruption processes, sufficient conditions for the positive recurrence of the single-server and multiserver queueing processes are established for the preemptive repeat different and the preemptive resume service disciplines.

Key words: Queueing theory, Server interruptions, Stability, Regenerative processes.

1. Introduction

For queueing systems with service interruptions [1], service to customers is regularly suspended. Service interruptions may result from resource sharing or server breakdowns and failures. In particular, when several queues share a common server, service interruptions are a natural abstraction to model the access of the other queues to the server. From the viewpoint of the customers of a particular queue, the server is unavailable whenever it serves customers from other queues.

Clearly, the service interruptions paradigm is most useful when the interruption process is independent of the arrival and service processes. If this is the case, the original queueing problem breaks down into two easier problems: the determination of the characteristics of the interruption process and the analysis of the queueing system with interruptions. In particular, such an approach greatly simplifies the analysis of queueing systems with a preemptive priority discipline. For the low priority customers of queueing systems with such a priority discipline, service interruptions occur when a high priority customer arrives during a low priority customer's service time. In accordance with the preemptive priority discipline, the low priority customer immediately leaves the server such that the high priority customer can enter the server upon arrival. After all high priority customers are served, the interrupted customer reenters the server and either continues its service (preemptive resume), restarts its service (preemptive repeat) or restarts its service with a new service time (preemptive repeat different). If the arrival and service processes of the high and low priority customers are independent, it is easily verified that the arrival, service and interruption processes of the low priority queue are independent as well.

Note that for some applications, independence of the interruption process cannot be assumed. For example, in a preemptive-repeat first-come-last-served queueing system, zero-length interruptions can be

*Corresponding author
[1]The work of the author is supported by Russian Foundation for Basic Research under grant 10-07-00017.
[2]The author is a post-doctoral fellow with the Research Foundation, Flanders (F.W.O.-Vlaanderen).

introduced to model the service interruptions upon arrival of new customers. However, in this case synchronisation of arrival and interruption process is required and the interruption process and queueing process cannot be studied separately.

White and Christie [2] were the first to study queueing systems with interruptions in connection with priority queueing systems. These authors investigated the $M/M/1$ queueing system with a preemptive resume priority discipline. Their results were later extended by Avi-Itzhak and Naor [3] and Thiruvengadam [4] who study preemptive resume priority queues with general service times. Gaver Jr. investigated the preemptive repeat and the preemptive repeat different priority disciplines [5]. Queues with batch Poisson arrivals and generally distributed service times were studied. Nain [6] has used the interruption paradigm to retrieve diffusion approximations for preemptive resume priority queues. More recently, Fiems, Steyaert and Bruneel [7] consider the discrete-time $M^X/G/1$ queueing system with a preemptive resume, preemptive repeat and a preemptive partial repeat priority discipline. These authors also provide expressions for the generating functions of the idle and busy periods of the queueing system with interruptions which enables them to study multi-class preemptive priority systems.

Apart from a priority scheduling discipline, a single server can be shared by multiple queues by polling or a cyclic service discipline. The server cyclically visits the different queues and remains with a particular queue until this queue is completely empty (exhaustive service) or until all customers are served that were present upon arrival of the server at the queue (gated service). Additionally, a limit can be placed upon the number of customers that are served during a visit (number-limited) or upon the duration of the visit (time-limited). For the exhaustive and the gated polling disciplines, the queueing process relates to a branching process and therefore exact expressions are available for the various performance measures of interest [8, 9]. For (exhaustive) number- and time-limited polling systems no exact results are available and various approximation methods have been proposed. In particular, the decomposition method focuses on a single queue of the polling system and models the access of the other queues as server interruptions. However, the arrivals, service and interruption processes are no longer independent for these systems, even if the arrival and service processes of the different queues are independent. An iterative procedure is then devised to find the stochastic characteristics of the interruption process.

The literature on limited polling systems is mostly concerned with number-limited [10, 11] and non-preemptive time-limited polling systems [12, 13, 14]. Only few authors consider preemptive time-limited polling systems which are the most interesting from the vantage point of the present contribution since service interruptions are allowed. De Souza e Silvia et al. [15] investigate such a polling system with Poisson arrivals, exponential service times, finite capacity buffers and a preemptive resume polling discipline. In this particular traffic setting, the polling discipline coincides with the preemptive repeat different discipline due to the lack of memory of the exponential service time distribution. Frigui and Alfa [16] focus on a time-limited polling system with a preemptive repeat different polling discipline as an approximation for the preemptive resume polling discipline. These authors assume a Markovian arrival process, phase-type service times and finite capacity buffers.

As already mentioned, interruptions also result from server failures or breakdowns. Some of the authors of the interruption models for priority queues discussed above [2, 3, 4, 5, 6, 7] also consider the case where the interruptions are triggered by server breakdowns. In addition, various authors also consider server breakdowns outside the priority queueing context. Federgruen and Green [17] provide bounds and approximations for the $M/G/1$ queue with generally distributed on- and off-times and a preemptive resume discipline. Notice that we adopt the priority queueing terminology to indicate how service is taken up after the interruption. Generally distributed on- and off-periods are also considered by Bruneel [18] for discrete-time queueing systems. Here, single-slot service times are considered such that there is no service preemption. Also Lee [19] considers a discrete-time queueing system with single slot service times but investigates a Markovian interruption process. Discrete-time queueing systems with generally distributed service times are investigated by Fiems et al. [20, 21]. In these papers, some generalisations of preemptive-repeat service disciplines are investigated. Tang [22] considers Poisson breakdowns when the server is working and renewal type breakdowns when it is idle whereas Li, Shi and Chau [23] investigate the transient behaviour of the $M/G/1$ queue subject to Poisson breakdowns. In both contributions a preemptive resume discipline is adopted. This is also the case in [24] where Nuñez Queija considers a processor sharing queue with Poisson

breakdowns. More recently, Balcioğlu et al. [25] approximate a $GI/D/1$ queue with correlated server breakdowns and preemptive resume by the corresponding system with an interruption process with (independent) hyper-exponentially distributed on-times and general off-times. Fiems et al. [26] study the $M/G/1$ queue where the server is simultaneously subjected to preemptive resume breakdowns and either preemptive repeat different or preemptive repeat identical breakdowns. Multiple server queues with breakdowns are studied by Mitrany and Avi-Itzhak [27] and Neuts and Lucantoni [28]. Both contributions consider a Poisson arrival and breakdown process and exponential service times. In the former contribution, server repair starts immediately and the repair times are exponentially distributed. In the latter contribution, the servers are repaired only when a number of servers have broken down.

Although queueing systems with service interruptions have been investigated since the late 1950s, to the best of our knowledge, stability conditions for these systems have not yet been formally proved. Such conditions are not straightforward though; typically the queueing systems are not work-conserving and therefore not always stable when the server capacity exceeds the arrival load. For preemptive repeat disciplines one intuitively sees that some service capacity is lost whenever service is interrupted. This contribution is the first to tackle the problem of the stability of queues with service interruptions with a regenerative stochastic process approach [33, 34], enabling us to prove the obtained conditions rigorously. The main contribution of the current work is its generality. The stability conditions we present below are valid for multiserver queues with generally distributed inter-arrival and service times and generally distributed server on- and off-times. Further, both preemptive repeat different and preemptive resume service disciplines are investigated. The conditions we obtain are simple and general, and therefore easily applicable. An important advantage of the presented approach is that it applies to non-Markovian processes as can be seen from the analysis below.

Before proceeding, we mention some applications of the queueing model at hand. In line with the literature described above, the server interruptions can be used to model high-priority traffic. In particular, stability of the low priority queue in a multiserver queueing system can be investigated, where each server has its dedicated stream of high-priority traffic. In addition, the server interruptions can model production interruptions in a production system with parallel possibly non-identical servers, each server being prone to errors. Interruptions then correspond to the server repair times and production may (preemptive repeat) or may not (preemptive resume) be lost if an error occurs. The stability results established in this paper then correspond to the maximum load that the respective buffer systems can sustain.

Key elements of our analysis are the synchronisation (coupling) of renewal processes based on the discrete structure of the distributions and a characterisation of the limit forward renewal process of the regenerations. Moreover, we use a new approach to extend the stability analysis to general initial states of the basic regenerative process. An important motivation for the paper was also to present different techniques in the framework of regenerative method for a wide class of models. For this reason, a refinement of the stability condition is included as well.

For completeness, we mention the fluid approximation approach as an alternative to the regenerative approach followed here. The fluid approximation approach replaces the stochastic model by a deterministic analogue termed the fluid model and stability of the original system is deduced from the stability of this fluid model. Such an approach has lead to significant progress in stability analysis of multiclass queueing networks [29, 30, 31]. See also Foss and Konstantopoulos [32] for a survey of various approaches to stability of queueing systems with a focus on the fluid approach. Nevertheless, the present paper does not rely on a fluid approach since the regenerative method turns out to be suitable to obtain complete and transparent proofs as well as natural stability conditions for the model at hand.

The remainder of this contribution is organised as follows. In the next section the queueing system is introduced and notation is established. Section 3 then concerns the regenerative structure of the multiserver queueing process. Stability conditions are established in section 4 and further extended in Section 5. In particular, the latter section concerns the inclusion of arbitrary initial states. Section 6 is devoted to some refinements of the main stability conditions for the single-server queueing system. Finally, conclusions are drawn in section 7.

2. Queueing system and notation

We consider a discrete time queueing system. Time is divided into fixed length intervals or slots and all arrivals and departures are synchronised with respect to slot boundaries. Therefore, service times and interarrival times are expressed in terms of numbers of slots. We here assume that the service times of the consecutive customers constitute a sequence of independent and identically distributed (iid) non-negative random variables; let S_n denote the service time of the nth customer. Similarly, the interarrival times between the consecutive customers also constitute a sequence of iid non-negative random variables; let τ_n denote the interarrival time between the nth and the $(n+1)$st customer. For further use, the arrival instant of customer $n+1$ is denoted by $A_n = \tau_1 + \cdots + \tau_n,\ n \geq 1$, $A_0 = 0$, whereas the residual renewal time of the arrival process $A = \{A_n, n \geq 0\}$ is given by,

$$A(t) = \min_k \{A_k - t : A_k - t \geq 0\}, \quad t \in \mathbb{N} = \{0, 1, 2, \ldots\}\,.$$

There are m servers which are unavailable from time to time. Let $X_n^{(i)}$ and $Y_n^{(i)}$ denote the length of the nth blocked and nth available period of the ith server, respectively $(i = 1, \ldots, m)$, whereby it is assumed that the first blocked period starts at slot boundary 0 for each server. In other words, all servers are down at time 0. The consecutive $(X_n^{(i)}, Y_n^{(i)})$ constitute sequences of iid random variables for all $i = 1, \ldots, m$. However, for each n and i, $X_n^{(i)}$ and $Y_n^{(i)}$ are not assumed to be independent. Let $Z_n^{(i)} = X_n^{(i)} + Y_n^{(i)}$ denote the length of the nth cycle of server i; a cycle consists of a blocked period followed by an available period, $i = 1, \ldots, m$. Since, for each i, $\{Z_n^{(i)}, n \geq 1\}$ is a sequence of iid random variables, we introduce the (zero-delayed) renewal processes $T^{(i)} = \{T_n^{(i)},\ n \geq 1\}$,

$$T_n^{(i)} = Z_1^{(i)} + \cdots + Z_n^{(i)}\,, \quad n \geq 1\,, \quad T_0^{(i)} = 0\,,$$

as well as the corresponding forward renewal time processes,

$$T_i(t) = \min_k \{T_k^{(i)} - t : T_k^{(i)} - t \geq 0\},\ \ i = 1, \ldots, m; \quad t \in \mathbb{N}\,.$$

In addition to the random variables introduced above and throughout the paper, indices of sequences of iid random variables are suppressed to denote generic elements of these sequences; for example, $X^{(i)}$ denotes a generic blocked period of server i, τ denotes a generic inter-arrival time, etc.

A server is free if it is neither serving customers nor interrupted. If a server becomes free and there are customers in the queue, a new customer enters the server. It is possible that more than one server becomes free simultaneously. Customers in the queue then choose a free server according to some algorithm, possibly random, for which no further restrictions are imposed.

The combination of multiple-slot service times and (independent) service interruptions, implies that an unavailability period can start while a customer is receiving service; service is then immediately interrupted. We here adopt either the preemptive resume service discipline or the preemptive repeat different service discipline. In the former case, service continues after the interruption whereas service is repeated from the start in the latter case. The service time after the interruption is independent of the original service time. In both cases, customers remain with the same server until service completion. Note that there are some technical complications if customers are allowed to change server. Intuitively, changing servers seems the right thing to do if another server is available when an interruption starts. However, in this case, one should avoid that customers constantly join servers that do not remain available for a sufficiently long time (such that their service is interrupted over and over again). Moreover, there may be technical reasons which forces customers to stay at a particular server.

Remark 1. The dependence between X_n and Y_n is natural in the context of GI/G/1 preemptive priority queues. Let $\hat{S}_n$ and $\hat{\tau}_n$ denote the service time of the nth high-priority customer and the interarrival time between the nth and the $(n+1)$st high-priority customer, respectively. From the vantage point of a low-priority customer, service is available whenever there are no high-priority customers in the system. The

low-priority queue can thus be modelled as a queueing system with interruptions whereby interruption and available periods correspond to busy and idle periods of the high priority queue, respectively. Hence, the lengths of the consecutive interruption and available periods can be expressed as follows,

$$X_n^{(1)} = \sum_{i=Q_n}^{Q_{n+1}-1} \hat{S}_i\,, \quad Y_n^{(1)} = \sum_{i=Q_n}^{Q_{n+1}-1} (\hat{\tau}_i - \hat{S}_i),$$

with customer Q_n initiating the nth busy period of the high-priority class,

$$Q_0 = 0,\ Q_n = \inf\{k > Q_{n-1} : \sum_{i=Q_{n-1}}^{k} (\hat{S}_i - \hat{\tau}_i) < 0\},\ n \geq 1\,.$$

Clearly, if both sequences $\hat{S}_n$ and $\hat{\tau}_n$ are iid, the sequence $(X_n^{(1)}, Y_n^{(1)})$ is iid too, although $X_n^{(1)}$ and $Y_n^{(1)}$ are dependent.

For queues with more than 2 classes, consecutive idle and busy periods are no longer independent for generally distributed inter-arrival times since the arrival processes of higher-priority customers do not regenerate at the same epoch. Therefore we have limited the discussion to two classes.

3. Regenerative structure of the queueing process

We have $m+1$ independent renewal processes A and $T^{(i)}$, $i = 1, 2, \ldots, m$. For the construction of regenerations, these processes have to be synchronised in such a way that common renewal points are obtained. It is well-known that in continuous time a synchronisation of two processes can be achieved by splitting and coupling under a regularity property of the densities of the involved renewal processes. Namely, at least one of these renewal-time distributions must be *spread-out*, that is a convolution of this distribution with itself has a density with respect to the Lebesgue measure. It is also possible to synchronise any number of independent renewal inputs under suitable assumptions; for the definition and more details see [35, 36].

However, it is not enough to construct a common renewal point for the superposed process to obtain regeneration of the queueing process (the queue size or workload process). The main difficulty in the continuous-time setting lies in the construction of a common renewal point such that the queue process simultaneously achieves *zero state* (or any other regeneration state, see below). The difficulty comes from the construction of a common renewal point, which is based on *splitting and coupling*. In typical situations a change of the original distributions makes it difficult to synchronise a common renewal point with a fixed state of the queueing process.

Fortunately, the situation is more tractable in the discrete-time setting. In particular, construction of common renewal points can be achieved without extra assumptions which contrasts with the *spread-outness* required in continuous time. Indeed, denote $\mathsf{P}(Z^{(i)} = k) = p_k^{(i)}$ and $\mathsf{P}(\tau = k) = q_k$. Since (by the natural assumption) cycle lengths and interarrival times are proper random variables,

$$\mathsf{P}(Z^{(i)} < \infty) = \sum_{k\geq 1} p_k^{(i)} = \mathsf{P}(\tau < \infty) = \sum_{k\geq 1} q_k = 1\,.$$

Therefore, for each i, there exist constants k_i and n_0 such that $p_{k_i} > 0$ and $q_{n_0} > 0$, $i = 1, \ldots, m$. Hence the processes have the same renewal interval $\kappa = n_0 \prod_{i=1}^m k_i$ with positive probability $q_{n_0}^{\kappa/n_0} \prod_{i=1}^m p_{k_i}^{\kappa/k_i} > 0$. For further use, we introduce the renewal process,

$$\gamma_0 = 0,\ \gamma_{n+1} = \inf(t > \gamma_n : T_i(t) = A(t) = 0,\ i = 1, \ldots, m),\ n \geq 0, \tag{1}$$

describing the common renewal points of the superposed process $A \cup \bigcup_i T^{(i)}$, and let $\gamma(t)$ denote the residual renewal time of this process at instant t,

$$\gamma(t) = \inf_{k>0}(\gamma_k - t : \gamma_k - t \geq 0),\ t \geq 0\,. \tag{2}$$

Note that $\gamma(0) = \gamma_1$.

In contrast to synchronisation of arrival and interruption processes, synchronisation with a fixed state of the queueing process however requires additional assumptions as shown further. We now construct a family of classical regenerations of the queue-size process under the assumption (for the moment) that the synchronisation mentioned above exists. Let $\nu(t)$ denote the number of customers in the system (in the queue and in the servers) at instant t; $\nu(t)$ excludes the departures at instant t but includes arrivals at that instant; we thus observe after possible departures and arrivals.

Define $\beta_0 = 0$ and

$$\beta_{n+1} = \inf_{t \geq 1}(t > \beta_n : T_1(t) = \cdots = T_m(t) = A(t) = 0,\ \nu(t) = 1),\ n \in \mathbb{N}. \tag{3}$$

As usual, it is assumed that $\inf \emptyset = \infty$. It is easy to see that the instants β_n constitute a renewal process of classical regenerations of the process $U := \{(\nu(t), T_1(t), \cdots, T_m(t), A(t)),\ t \geq 0\}$.

Our goal is to find assumptions which guarantee the existence of the renewal process $\beta = \{\beta_n\}$ with finite mean regeneration period. The latter property is called *positive recurrence*. More exactly, we call the renewal process (3) positive recurrent if

$$\beta_1 < \infty \quad \text{with probability 1 and } \mathsf{E}\{\beta_2 - \beta_1\} := \alpha_0 < \infty. \tag{4}$$

We characterise the recurrence property of the renewal process β by the limiting behaviour of the forward regeneration time at instant t, which is defined as follows:

$$\beta(t) = \inf_k(\beta_k - t : \beta_k - t \geq 0),\ t \in \mathbb{N}. \tag{5}$$

Hence, for any instant t, the first joint renewal epoch of the arrival process and all interruption processes from time t onwards, occurs at time $t + \beta(t)$. It is known [37, pp. 366] that

$$\alpha_0 = \infty \text{ implies } \quad \beta(t) \Rightarrow \infty \quad \text{(in probability)}, \tag{6}$$

regardless of the initial value $\beta(0)$. The key step of the stability analysis presented below is to establish that $\beta(t) \not\Rightarrow \infty$ (in probability) which implies $\alpha_0 < \infty$. Such an approach has been successfully applied before, see amongst others [34, 38].

In the remainder, we consider the zero-delayed process β first. In this case, we have $T_1(0) = \cdots = T_m(0) = A(0) = 0$ and $\nu(0) = 1$ such that the following stochastic equivalence holds,

$$\beta_1 =_{st} \beta_2 - \beta_1.$$

and such that $\mathsf{E}\beta_1 = \alpha_0$. Then $\alpha_0 < \infty$ implies $\beta_1 < \infty$ with probability 1 (w.p.1), and positive recurrence (4) follows.

For the zero-delayed process, we call the process U positive recurrent if the renewal process β is positive recurrent. Probabilities and expectations are indicated by P and E, respectively. For a more general initial state $U(0) = z$, probabilities and expectations are indicated by P_z and E_z as usual. However, for a general initial state z, β is delayed. This means that, in general, the first regeneration period β_1 has a different distribution, and stability analysis requires some extra effort to establish finiteness of this first regeneration period, $\mathsf{P}_z(\beta_1 < \infty) = 1$. In Section 5, we present a new approach to address this issue.

Remark 2. The process U also regenerates when $T_1(t) = \cdots = T_m(t) = A(t) = 0$ and $\nu(t) = i$ for all $i \in \mathbb{N}^+ = \{1, 2, \ldots\}$. By the interruption strategy adopted, the service times of customers that are interrupted at time t are resampled which implies that U regenerates. Nevertheless, in the remainder we solely focus on the regeneration instants $\{\beta_n\}$ defined above.

Arrival process	
τ_n	interarrival time between the nth and $(n+1)$st customer
A_n	arrival instant of customer $n+1$
$A(t)$	residual renewal time of the arrival process
Interruption process	
$X_n^{(i)}$	length of the nth blocked period of server i
$Y_n^{(i)}$	length of the nth available period of server i
$Z_n^{(i)}$	length of the nth cycle of server i
$T_n^{(i)}$	completion time of the nth cycle of server i
$T_i(t)$	residual renewal time of the cycle process of server i
Other processes	
$\nu(t)$	queue content at time t
β_n	nth renewal instant of the complete process
$\beta(t)$	residual renewal time of the complete process
γ_n	nth renewal instant of the superimposed process of the arrivals and interruptions
$\gamma(t)$	residual renewal time of the superimposed process of the arrivals and interruptions

Table 1: Summary of notation used throughout the paper.

4. Main stability results

We now present the basic stability result. To facilitate the exposition, table 1 summarises the notation introduced in Sections 2 and 3. We first consider the multiserver queue with a preemptive repeat interruption discipline. Afterwards, the preemptive resume discipline is investigated. In either case, the zero-delayed process is considered: the arrival process and the cycle processes regenerate at $t = 0$. At instant $t = 0$ there is a single customer in the queue which arrived at that instant.

4.1. Preemptive repeat

We consider a queueing system with m servers and recall that superscripts are used to distinguish between the random processes related to the different servers. Let $N(t) = \min(k \geq 1 : A_k \geq t)$ be the number of arrivals in interval $[0, t]$, $t \geq 0$. Moreover, we assume that,

$$0 < \lambda := \frac{1}{\mathsf{E}\tau} < \infty, \quad 0 < \lambda_0^{(i)} := \frac{1}{\mathsf{E}Z^{(i)}} < \infty, \quad \lambda_0 = \sum_{1 \leq i \leq m} \lambda_0^{(i)}. \tag{7}$$

Here, λ denotes the arrival rate and $\lambda_0^{(i)}$ denotes the cycle rate of server i. The cycle rate is the inverse of the mean cycle time. Also some assumptions are introduced which ensure that synchronisation is achieved. In particular, assume the existence of integer $\theta_1, \ldots, \theta_m$ such that

$$\prod_{i=1}^{m} \mathsf{P}(\tau = \theta_i Z^{(i)},\, Y^{(i)} > S) > 0. \tag{8}$$

Note that a wide class of discrete distributions satisfy assumption (8). This assumption is for example automatically satisfied if $Y^{(i)}$ has infinite support (for $i = 1, \ldots, m$) and τ's support equals the non-negative integers. In this particular case, there is no need to impose further restrictions on the service time distribution. To illustrate this assumption by a concrete example, let $m = 2$, interarrival time τ follow a Poisson distribution and $Y^{(i)}$ be unbounded (and independent of $X^{(i)}$), $i = 1, 2$. Then, because $\mathsf{E}S < \infty$, it is easy to check that (8) holds with $\theta_1 = \theta_2 = 1$.

We now have the following theorem.

Theorem 1. *Assume that conditions (7) and (8) hold and that the interarrival times τ and cycles $Z^{(i)}$, $i = 1, \ldots, m$, are aperiodic. If the following negative drift condition is satisfied,*

$$(\lambda + \lambda_0)\mathsf{E}S + \sum_{i=1}^{m} \lambda_0^{(i)} \mathsf{E}X^{(i)} < m\,, \tag{9}$$

then the queue size process $\{\nu(t),\, t \geq 0\}$ is positive recurrent with respect to the zero delayed renewal process β as defined in (3).

Proof. Let

$$\mu(t) = \sum_{n=1}^{t} I(\nu(n) < m)$$

be the total time when the queue content is less than the number of servers, within interval $[0,\, t]$. Further, let $\mu_0^{(i)}(t)$ and $X_i(t)$ denote the idle time and the blocked time of server i in the interval $[0, t]$ respectively, and let $\mu_0(t)$ and $X(t)$ denote the total idle time and the total blocked time of all servers during this interval:

$$\mu_0(t) = \sum_{i=1}^{m} \mu_0^{(i)}(t)\,, \quad X(t) = \sum_{i=1}^{m} X_i(t)\,.$$

Since $\mu(t)$ counts the slots where at least one server is idle and $\mu_0^{(i)}(t)$ counts the slots where server i is idle, we obviously have,

$$\mu(t) \geq \mu_0^{(i)}(t) \quad \text{for all } i,\, t \geq 0. \tag{10}$$

Let $W(t)$ denote the workload — the sum of the (remaining) service time of all customers in the queue — at time t. In view of the interruption discipline, it is assumed that whenever the service of a customer is interrupted, the remaining service time of this customer is served immediately and a new customer service time is added to the workload.

Since $\nu(0) = 1$, we have $W(0) < \infty$ w.p.1. Now we have the following lower bound for the arrived workload $V(t)$ within interval $[0,\, t]$:

$$\begin{aligned} V(t) &\geq W(t) + \sum_{i=1}^{m} (t - \mu_0^{(i)}(t) - X_i(t)) \\ &\geq mt - \mu_0(t) - X(t) \\ &\geq mt - m\mu(t) - X(t)\,. \end{aligned}$$

$V(t)$ includes repeated service as well as the initial workload $V(0) = W(0)$. For the first inequality, the remaining service time of interrupted service is neglected. The second inequality neglects the workload at instant t whereas the third inequality follows from (10). Summarising, we find,

$$m\mu(t) \geq mt - V(t) - X(t). \tag{11}$$

Let $F_i(t)$ denote the number of interrupted service times of server i in interval $[0, t]$ and $G_i(t)$ denote the number of breakdowns of server i starting within the interval $[0, t]$. To simplify notation, we further assume that the service times which are assigned after interruptions are selected from a doubly indexed sequence of the iid random variables $S_j^{(i)}$, distributed as S, where $S_j^{(i)}$ is the service time assigned after the jth interruption at server $i = 1, \ldots, m$, $j \geq 1$. Since there is at most one service interruption for every blocked period, we find,

$$V(t) = W(0) + \sum_{j=1}^{N(t)} S_j + \sum_{i=1}^{m} \sum_{j=1}^{F_i(t)} S_j^{(i)} \leq \sum_{j=1}^{N(t)} S_j + \sum_{i=1}^{m} \sum_{j=1}^{G_i(t)} S_j^{(i)}. \tag{12}$$

By the strong law of large numbers (SLLN) for renewal processes, we have w.p.1 as $t \to \infty$,

$$\frac{G_i(t)}{t} \to \lambda_0^{(i)}, \quad \frac{X_i(t)}{t} \to \lambda_0^{(i)} \mathsf{E} X^{(i)}, \quad \text{for } i = 1, \ldots, m,$$
$$\frac{1}{t}\Big(\sum_{j=1}^{N(t)} S_j + \sum_{i=1}^{m} \sum_{j=1}^{G_i(t)} S_j^{(i)}\Big) \to \lambda \mathsf{E} S + \sum_{i=1}^{m} \lambda_0^{(i)} \mathsf{E} S = (\lambda + \lambda_0) \mathsf{E} S\,. \tag{13}$$

Now (11), (12) and (13) imply

$$\liminf_{t\to\infty} m \frac{\mu(t)}{t} \geq m - (\lambda + \lambda_0)\mathsf{E} S - \sum_{i=1}^{m} \lambda_0^{(i)} \mathsf{E} X^{(i)}, \tag{14}$$

or,

$$\liminf_{t\to\infty} \frac{\mu(t)}{t} \geq 1 - \frac{(\lambda + \lambda_0)\mathsf{E} S}{m} - \frac{\sum_{i=1}^{m} \lambda_0^{(i)} \mathsf{E} X^{(i)}}{m} := \varepsilon_0 > 0\,. \tag{15}$$

The right-hand side of the last inequality is positive by the negative drift condition (9). Since $\mu(t)/t \geq 0$, Fatou's lemma and equation (15) imply,

$$\liminf_{t\to\infty} \frac{\mathsf{E}\mu(t)}{t} > 0\,.$$

Moreover, $\mathsf{E}\mu(t) = \sum_{1 \leq n \leq t} \mathsf{P}(\nu(n) < m)$, $t \geq 1$ such that the former inequality implies,

$$\mathsf{P}(\nu(t) < m) \not\to 0, \text{ as } t \to \infty.$$

This means that there exists a non-random (sub)sequence of time instants $n_k \to \infty$ (as $k \to \infty$) and some $\varepsilon > 0$ such that

$$\inf_k \mathsf{P}(\nu(n_k) < m) \geq \varepsilon\,. \tag{16}$$

All renewal periods have a finite mean value and the renewal intervals are aperiodic. Therefore, the following weak limits exist,

$$\lim_{t\to\infty} \mathsf{P}(T_i(t) = k) = \frac{1}{\mathsf{E} Z^{(i)}} \mathsf{P}(Z^{(i)} \geq k+1), \qquad \text{for } k \geq 0;\ i = 1, \ldots, m,$$
$$\lim_{t\to\infty} \mathsf{P}(A(t) = k) = \frac{1}{\mathsf{E}\tau} \mathsf{P}(\tau \geq k+1), \qquad \text{for } k \geq 0. \tag{17}$$

These limits hold for arbitrary initial states $T_i(0)$, $A(0)$, and in particular for $T_i(0) = A(0) = 0$. Hence, we have,

$$\lim_{t\to\infty} \mathsf{P}(T_i(t) = 0) = \lambda_0^{(i)},\ i = 1, \ldots, m; \quad \lim_{t\to\infty} \mathsf{P}(A(t) = 0) = \lambda.$$

Therefore, denote $\epsilon = \min_{0 \leq i \leq m}\{\lambda, \lambda_0^{(i)}\} > 0$. By (17), there exists a constant t_0 such that for all $t \geq t_0$, we have,

$$\mathsf{P}(T_i(t) = 0) \geq \frac{\epsilon}{2},\ i = 1, \ldots, m; \ \ \mathsf{P}(A(t) = 0) \geq \frac{\epsilon}{2}\,. \tag{18}$$

Recall that $\gamma(t)$ denotes the residual renewal time at instant t of the superimposed process $A \cup \bigcup_i T^{(i)}$; see equation (2). By (18) and by the independence of the renewal processes, we obtain that residual regeneration time $\gamma(t)$ satisfies,

$$\mathsf{P}(\gamma(t) = 0) = \mathsf{P}(T_i(t) = A(t) = 0,\, i = 1, \ldots, m) \geq \left(\frac{\epsilon}{2}\right)^{m+1} > 0\,, \quad t \geq t_0\,.$$

In other words, $\gamma(t) \not\Rightarrow \infty$ and positive recurrence of the (zero-delayed) process follows, $\mathsf{E}\gamma_1 < \infty$. In particular, the forward regeneration time process $\gamma(t)$, $t \geq 0$, is tight. We then conclude that there exists a constant D such that (see (16))

$$\inf_k \mathsf{P}(\nu(n_k) < m,\, \gamma(n_k) \leq D) \geq \frac{\varepsilon}{2}. \tag{19}$$

In the remainder of this proof, we focus on an arbitrary (fixed) n_k satisfying (16). Hence, by (19), a common renewal point ϕ_k (for all renewal processes) appears in the interval $[n_k,\, n_k + D]$ with positive probability $\geq \varepsilon/2$ and $\nu(\phi_k) \leq m + D$ since there are at most $m + D$ customers in the queue at this instant.

We rewrite assumption (8) as follows. There exist numbers j_0, b_i, u_i, $i = 1, \ldots, m$ and r_0 such that

$$\begin{aligned} &\mathsf{P}(\tau = j_0) > 0, \\ &\mathsf{P}(Z^{(i)} = b_i,\, Y^{(i)} = u_i) > 0,\, i = 1, \ldots, m, \\ &\mathsf{P}(S = r_0) = v_{r_0} > 0, \end{aligned} \tag{20}$$

and which are connected by $j_0 = b_i\theta_i$, $r_0 < u_i$, $i = 1, \ldots, m$. The conditions above allow us to unload the system with positive probability in a finite time while retaining synchronisation. Indeed, we realise (i) θ_i cycles of b_i slots of server i with u_i slots for the active periods, $i = 1, \ldots, m$; (ii) service times $S = r_0$ for all customers (being in the system at the instant ϕ_k and the new ones); (iii) interarrival times of j_0 slots. It then follows that (starting at the instant γ_{ϕ_k}) the number of customers being in the system at the beginning of a cycle is reduced at least by one as long as at least two servers are busy. Indeed even in that *worst* case the number of available periods included in the cycles is $\geq \min_{i \neq j}(\theta_i + \theta_j) \geq 2$ and thus the number of served customers within a cycle is not less than 2. Hence, we can unload the queue till there is a single customer in the queue and a regeneration occurs. Note that we realise the events $\{\tau = j_0\}$, at most for $(m + D - 1)$ arrivals, until a customer arrives in an empty queue (assuming that the event $\{\nu(n_k) < m,\, \gamma(n_k) \leq D\}$ holds). Such a scenario occurs in interval $[\phi_k,\, \phi_k + j_0(D + m - 1)]$ with a positive probability bounded below by,

$$[\mathsf{P}(\tau = j_0)]^{m+D-1} \prod_{i=1}^{m} [\mathsf{P}(Z^{(i)} = b_i,\, Y^{(i)} = u_i)]^{m+D-1} v_{r_0}^{2(m+D-1)} := \delta_0 > 0\,. \tag{21}$$

We conclude that on the event $\{\nu(n_k) < m,\, \gamma(n_k) \leq D\}$ a regeneration occurs in interval $[n_k,\, n_k + H]$ with a probability $\geq \varepsilon\delta_0/2$ whereby the length $H := D + j_0(D + m - 1)$ does not depend on n_k. In other words, the forward regeneration time at instant n_k satisfies

$$\mathsf{P}(\beta(n_k) \leq H) \geq \frac{\varepsilon}{2}\delta_0 > 0. \tag{22}$$

Since the lower bound is uniform in n_k, positive recurrence in the zero-delayed case follows. □

4.2. Preemptive resume

We now touch upon the stability of the model with preemptive resume service interruptions. As opposed to queues with preemptive repeat different interruptions, interrupted service continues when the server returns from a blocked period for queues with preemptive resume interruptions. It is intuitively clear that the negative drift condition must be changed to take into account both the arriving workload (as in the system without interruptions) and the blocked periods. However, for preemptive resume, the interruptions bring no additional workload.

Theorem 2. *The statement of Theorem 1 holds for the system with preemptive resume service interruptions if the negative drift assumption (9) is replaced by*

$$\sum_i \lambda_0^{(i)} \mathsf{E}X^{(i)} + \lambda \mathsf{E}S < m. \tag{23}$$

Proof. Indeed, in this case, $\mu(t)m \geq mt - X(t) - \sum_{i=1}^{N(t)} S_i$ and thus

$$\limsup_{t\to\infty} m\frac{\mu(t)}{t} \geq m - \sum_i \lambda_0^{(i)} \mathsf{E}X^{(i)} - \lambda \mathsf{E}S > 0. \tag{24}$$

This gives (38). The proof of the 2nd part of Theorem — appearance of a regeneration point in a finite interval — holds unchanged. □

Remark 3. The negative drift condition (9) has a nice qualitative explanation. The standard term $\lambda \mathsf{E}S$ relates to incoming workload as usual, the term $\sum_{i=1}^{m} \lambda_0^{(i)} \mathsf{E}X^{(i)}$ describes the loss of capacity caused by the interruptions, and finally, the term $\lambda_0 \mathsf{E}S$ expresses the loss caused by service repetitions of interrupted customers. To guarantee stability, it is enough to consider the behaviour of the process when the queue is heavily loaded. In this case the rate of the actual interruptions approaches the rate of the blocked periods because the queue is almost always busy. This effect has been observed in many other models, see for instance [34]. Nevertheless, the estimate of the capacity loss by the service repetitions can be further refined as shown in Section 6.2.

Although we do not include instability analysis in this paper, the proofs of the stability theorems suggest that the negative-drift condition (9) of Theorem 1 is tight in the sense that, for any given λ, λ_0, $\mathsf{E}X$ and $\mathsf{E}S$ that do not satisfy the negative-drift condition, distributions of the interarrival times, service times, available and unavailable periods can be found such that the system is not stable. However, this does not mean that the system is always unstable if (9) is not satisfied. Section 6.2 is concerned with tightening the negative-drift condition. However, the negative-drift condition found there is not an expression of a finite number of moments of the arrival, service and interruption processes.

5. Extension of the initial conditions

We now extend our stability result to the case of non-zero initial conditions. For this, we introduce the process $\hat{U} = \{\hat{U}(t),\, t \geq 0\}$ with $\hat{U}(t) = \{\nu(t),\, \gamma(t)\}$ and where $\gamma(t)$ is defined in (2). Note that the process $\hat{U}$ is not Markovian; we may extend the process $\hat{U}$ to a Markovian process if we include the residual service times $S^{(i)}(t)$ in the different servers, the remaining blocked times $X^{(i)}(t)$ and the remaining available times $Y^{(i)}(t)$ of the different servers and the remaining interarrival time $A(t)$ $(i = 1, \ldots, m)$. We may however restrict ourselves to the process $\hat{U}$ since the component $\gamma(t)$ dominates these processes, that is, $\gamma(t) \geq \max\{A(t),\, \gamma(t),\, S^{(i)}(t),\, X^{(i)}(t),\, Y^{(i)}(t),\, i = 1, \ldots, m\}$ w.p.1 for all t. Note that, as mentioned in the introduction, this shows an advantage of our approach: we do not need the Markov property.

Theorem 3. *Under the conditions of Theorem 1 and for any initial state $\hat{U}(0) = z := (z_1,\, z_2)$, the queue size process is positive recurrent with respect to the (delayed) renewal process β.*

Proof. We split the proof into two parts. First, we show that the time within the interval $[0,\, t]$ during which the basic process $\hat{U}$ is in a compact set increases to infinity as $t \to \infty$. In the second step, we show that the number of visits to the compact set by the process $\{\hat{U}(t),\, t \geq 0\}$ within the first regeneration period $[0,\, \beta_1)$ is finite w.p.1 for any initial state. From these facts, it immediately follows that the total number of regeneration cycles cannot be less than two, and thus, $\beta_1 < \infty$ w.p.1.

Part 1. Using the positive recurrence (and thus the tightness) of the process $\gamma(t)$, $t \geq 0$, one can find a constant D_0 such that

$$\lim_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t-1} I(\gamma(u) \leq D_0) \geq 1 - \frac{\varepsilon_0}{2} \quad \text{w.p.1}, \tag{25}$$

whereby ε_0 is defined in equation (15). The inequality above follows from the observation that the process $\sum_{u=0}^{t-1} I(\gamma(u) \leq D_0)$ is a cumulative process in terms of the positive recurrent process $\{\gamma(t)\}$; see [39]. By a slight adaptation of the argument leading to equation (15), we find,

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t-1} I(\nu(u) \geq m) = 1 - \liminf_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t-1} I(\nu(u) < m) \leq 1 - \varepsilon_0. \tag{26}$$

Introducing the compact set $\mathbf{B}_0 = [0,\, m) \times [0,\, D_0]$ and associated counting function $\Gamma_0(t) = \sum_{u=0}^{t-1} I(\hat{U}(u) \in \mathbf{B}_0)$, it then easily follows from (25) and (26) that

$$\liminf_{t\to\infty} \frac{\Gamma_0(t)}{t} \geq \lim_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t-1} I(\gamma(u) \leq D_0) - \limsup_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t-1} I(\nu(u) \geq m) \geq \varepsilon_0/2 > 0\,. \tag{27}$$

We conclude that the time within the interval $[0,\, t]$ during which the basic process $\hat{U}$ is in the compact set $\mathbf{B}_0$ increases to infinity as $t \to \infty$.

Part 2. Fix an arbitrary $\hat{U}(0) = z = (z_1,\, z_2)$ and take a bounded set $\tilde{\mathbf{B}}_0 = [0,\, B_1] \times [0,\, B_2]$ such that $z \in \tilde{\mathbf{B}}_0$ and $\mathbf{B}_0 \subseteq \tilde{\mathbf{B}}_0$. That is, choose $B_1 \geq \max(z_1,\, m-1)$ and $B_2 \geq \max(z_2,\, D_0)$. By a similar argument as the one leading to equation (22), it is easy to show that for $\hat{H} = B_2 + j_0(B_1 + B_2)$ and j_0 in accordance with equation (21), there exists a constant $\hat{\delta}_0 > 0$ such that,

$$\mathsf{P}_z(\beta(t) \leq \hat{H} | \hat{U}(t) = u,\, \beta_1 > t) := \mathsf{P}(\beta(t) \leq \hat{H} | \hat{U}(0) = z,\, \hat{U}(t) = u,\, \beta_1 > t) \geq \hat{\delta}_0\,, \tag{28}$$

for any $u \in \tilde{\mathbf{B}}_0$ and all $t \geq 0$. We then obtain

$$\begin{aligned}
1 &\geq \sum_{t \geq 0} \mathsf{P}_z(\hat{U}(t\hat{H}) \in \tilde{\mathbf{B}}_0,\, t\hat{H} < \beta_1 \leq (t+1)\hat{H}) \\
&= \sum_{t \geq 0} \mathsf{P}_z(\beta_1 \leq (t+1)\hat{H} | \hat{U}(t\hat{H}) \in \tilde{\mathbf{B}}_0,\, \beta_1 > t\hat{H}) \mathsf{P}_z(\hat{U}(t\hat{H}) \in \tilde{\mathbf{B}}_0,\, \beta_1 > t\hat{H}) \\
&\geq \hat{\delta}_0 \sum_{t \geq 0} \mathsf{P}_z(\hat{U}(t\hat{H}) \in \tilde{\mathbf{B}}_0\,, \beta_1 > t\hat{H})\,.
\end{aligned}$$

By analogy, we have for $n = 0, \ldots, \hat{H} - 1$,

$$1 \geq \hat{\delta}_0 \sum_{t \geq 0} \mathsf{P}_z(\hat{U}(t\hat{H} + n) \in \tilde{\mathbf{B}}_0,\, \beta_1 > t\hat{H} + n).$$

Summing up all inequalities, we obtain the following upper bound

$$\sum_{t \geq 0} \mathsf{P}_z(\hat{U}(t) \in \tilde{\mathbf{B}}_0,\, \beta_1 > t) = \mathsf{E}_z \left(\sum_{t=0}^{\beta_1 - 1} I(\hat{U}(t) \in \tilde{\mathbf{B}}_0) \right) \leq \frac{\hat{H}}{\hat{\delta}_0}. \tag{29}$$

Let $\tilde{G}_0 = \sum_{t=0}^{\beta_1 - 1} I(\hat{U}(t) \in \tilde{\mathbf{B}}_0)$ denote the number of visits to the set $\tilde{\mathbf{B}}_0$ by the process $\hat{U}$ in the first regeneration period $[0,\, \beta_1)$. The former equation then shows that $\mathsf{E}_z \tilde{G}_0 < \infty$ which in turn implies,

$$\mathsf{P}_z(\tilde{G}_0 < \infty) = 1. \tag{30}$$

In other words, the renewal process is in the compact set $\tilde{\mathbf{B}}_0$ during a finite number of slots during the first regeneration period. Since $\mathbf{B}_0 \subset \tilde{\mathbf{B}}_0$, the number of slots that the renewal process is in the compact set $\tilde{\mathbf{B}}_0$ grows unbounded and the statement of theorem follows. □

6. Stability of the single-server system: some refinements

In this section, we present some extensions of the stability result. In Section 6.1, we relax the aperiodicity assumption of the interarrival times which was imposed in the preceding section, thereby limiting ourselves to the single server case. As in the preceding sections, we first consider stability of the preemptive repeat discipline and then simplify our argument for the preemptive resume discipline.

The stability conditions of the interruption system are further refined as well. The conditions of the preceding section may fail to hold while the queueing process is positive recurrent. Tighter stability conditions are obtained in Section 6.2. For ease of notation and since there is only one server, we drop the superscripts which refer to the different servers.

6.1. Relaxing the aperiodicity assumption

Recall that τ and $Z = X + Y$ denote a generic interarrival time and a generic cycle length, respectively. Now we do not require aperiodicity of $Z^{(i)}$ and τ. To assure regeneration, we make the following assumptions,

$$\mathsf{P}(|\tau - Z| = 1) > 0\,, \tag{31}$$

and there exists some $k_0 \geq 1$ such that

$$\mathsf{P}\Big(Z = k_0\tau,\, Y > S_1 + \cdots + S_{k_0+1}\Big) > 0\,. \tag{32}$$

Assumption (31) is equivalent to the existence of numbers i_0 and l_0 such that

$$\max\Big(q_{i_0}p_{i_0+1},\, q_{l_0+1}p_{l_0}\Big) > 0\,. \tag{33}$$

Recall that p_k (q_k) denotes the probability that the cycle length (interarrival time) equals k slots. Moreover, assumption (32) can be reformulated in terms of the given distributions as follows. There exist numbers $j_0 > 0$ and $0 < u_0 < k_0 j_0$ such that

$$\mathsf{P}\Big(\tau = j_0\Big)\mathsf{P}\Big(Z = k_0 j_0,\, Y = u_0\Big)\mathsf{P}\Big(u_0 > S_1 + \cdots + S_{k_0+1}\Big) > 0\,. \tag{34}$$

Here $v_n = \mathsf{P}(S = n)$ denotes the probability that the service time equals n slots. Notice that $\mathsf{P}(u_0 > S_1 + \cdots + S_{k_0+1}) > 0$ if and only if there exists a number $r_0 > 0$ such that

$$v_{r_0} > 0 \text{ and } u_0 > (k_0 + 1)r_0. \tag{35}$$

Assumption (34) means that the number of customers served within the active period Y exceeds the number of new arrivals k_0 during period $Z = k_0\tau$ with a positive probability. Note that a wide class of discrete distributions satisfy assumptions (31) and (32).

Theorem 4. *Assume that conditions (31) and (32) hold and that,*

$$\lambda := \frac{1}{\mathsf{E}\tau} \in (0, \infty), \quad \lambda_0 := \frac{1}{\mathsf{E}Z} \in (0, \infty)\,. \tag{36}$$

Moreover, assume that the following negative drift condition is satisfied,

$$\lambda_0 \mathsf{E}X + (\lambda + \lambda_0)\mathsf{E}S < 1\,. \tag{37}$$

Then the zero-delayed queue-size process $\nu = \{\nu(t), t \geq 0\}$ is positive recurrent with respect to the zero-delayed process β.

Proof. As in Theorem 1, we use negative drift assumption (9) to establish that there exists a non-random (sub)sequence of time instants $n_k \to \infty$ (as $k \to \infty$) and some $\varepsilon > 0$ such that,

$$\inf_i \mathsf{P}(\nu(n_k) = 0) \geq \varepsilon\,. \tag{38}$$

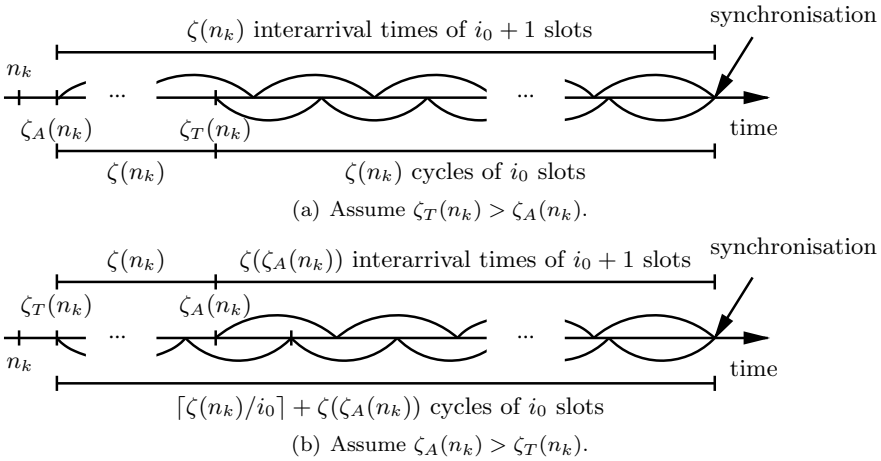
We now use the tightness of the residual renewal time processes $\{T(t),\, t \geq 0\}$ and $\{A(t),\, t \geq 0\}$. This well-known property (under finite mean interrenewal times) can be obtained from the weak convergence of the residual renewal time to a proper limit in the aperiodic case and also holds when the renewal interval is periodic, for more detail see [35, 33]. By the tightness, one can find a constant $D < \infty$ such that (38) implies,

$$\inf_i \mathsf{P}\Big(\nu(n_k) = 0,\, T(n_k) \leq D,\, A(n_k) \leq D\Big) \geq \frac{\varepsilon}{2}. \tag{39}$$

Let $\zeta_A(n) := n + A(n)$ and $\zeta_T(n) \equiv n + T(n)$ denote the first renewal after (or at) boundary n of the arrival and interruption process, respectively. Further, denote $\zeta(n) = |\zeta_T(n) - \zeta_A(n)|$. In the event

$$\mathcal{E}(n_k, D) = \Big\{\nu(n_k) = 0,\, T(n_k) \leq D,\, A(n_k) \leq D\Big\}\,,$$

we have $\zeta(n_k) \leq D$. Recall that assumption (31) implies the existence of numbers i_0 and l_0 such that $p_{i_0}q_{i_0} > 0$ and/or $p_{l_0+1}q_{l_0} > 0$. We now consider the two cases separately, assuming the event $\mathcal{E}(n_k, D)$ holds.



(a) Assume $\zeta_T(n_k) > \zeta_A(n_k)$.

(b) Assume $\zeta_A(n_k) > \zeta_T(n_k)$.

Figure 1: Synchronisation of arrival and interruption processes.

Case 1: $p_{i_0}q_{i_0+1} > 0$. First assume $\zeta_A(n_k) \leq \zeta_T(n_k)$. In this case, we realise $\zeta(n_k)$ interarrival times of length i_0+1 and $\zeta(n_k)$ cycles of length i_0 such that a common renewal point is reached at instant

$$\zeta_A(n_k) + (i_0+1)\zeta(n_k) \leq n_k + (i_0+1)D$$

with probability $\geq (q_{i_0+1}p_{i_0})^D > 0$; see figure 1(a). By construction, it is further observed that there are $\zeta(n_k) \leq D$ customers in the queue at this common renewal instant.

Now assume $\zeta_A(n_k) > \zeta_T(n_k)$. We realise cycles of the interruption process of length i_0 until the renewal instant exceeds or equals $\zeta_A(n_k)$ as shown in figure 1(b). Clearly, $\lceil\zeta(n_k)/i_0\rceil$ such cycles are required which are realised with probability $\geq p_{i_0}^{\lceil D/i_0\rceil} > 0$. By realising $\zeta(\zeta_A(n_k))$ cycles of length i_0 and $\zeta(\zeta_A(n_k))$ interarrival times of length i_0+1, a common renewal point is reached at instant

$$\zeta_A(n_k) + \zeta(\zeta_A(n_k))(i_0+1) \leq n_k + D + i_0(i_0+1)$$

with probability $\geq p_{i_0}^{\lceil D/i_0\rceil+i_0}q_{i_0+1}^{i_0} > 0$. Here, we used the fact that the *overshoot* $\zeta(\zeta_A(n_k))$ of the interruption process at instant $\zeta_A(n_k)$ is bounded by i_0. By construction, there are at most $\zeta(\zeta_A(n_k)) \leq i_0$ customers in the queue at the common renewal instant.

Hence, we find that for each n_k a common renewal point of A and T can be realised within a finite interval $[n_k, n_k + \max(D(i_0+1), D + i_0(i_0+1))]$ with probability $\geq$

$$\min\Big((q_{i_0+1}p_{i_0})^D,\, p_{i_0}^{\lceil D/i_0\rceil+i_0}q_{i_0+1}^{i_0}\Big) > 0$$

such that there are at most $\max(i_0, D)$ customers in the queue at this instant. By (34) we can then realise a cycle of length k_0j_0 during which at least k_0+1 customers are served and k_0 interarrival times of length j_0 are realised with positive probability. We can thus reduce the queue size while retaining synchronisation between T and A. At most $\max(i_0, D) - 1$ such realisations are required.

Summarising, we find a regeneration point with positive probability in the interval

$$\Big[n_k, n_k + \max(D(i_0+1),\, D + i_0(i_0+1)) + (\max(i_0, D) - 1)k_0j_0\Big].$$

Neither the length of the interval nor the probability depend on n_k. Because the sequence $\{n_k\}$ is non-random, positive recurrence (4) follows by the characterisation (6).

Case 2: $p_{l_0+1}q_{l_0} > 0$. With a minor modification of the previous considerations, one can again construct a common renewal point of T and A within a finite interval

$$\Big[n_k, n_k + \max(D(l_0+1),\, D + l_0(l_0+1))\Big]$$

with probability $\geq$

$$\min\Big((p_{l_0+1}q_{l_0})^D,\, p_{l_0+1}^{l_0} q_{l_0}^{\lceil D/l_0\rceil + l_0}\Big) > 0.$$

The queue size at this common renewal point is bounded by $\max(\lceil D/l_0\rceil + l_0, D)$ and can be reduced by realising cycles, service times and interarrival times as in case 1. Finally, the positive recurrence (4) follows by the characterisation (6). □

Remark 4. We believe that the most restrictive assumptions (8) and (34) can be replaced by some other assumptions. But we do not think any such assumptions may be (much) less restrictive because, to some extent, it is the price that needs to be payed for the complicated and delicate coupling procedure used in the proofs of Theorem 1 and 4.

6.2. Tighter stability conditions

As noted in Remark 3, the negative-drift condition (9) of Theorem 1 is tight in the sense that, for any given λ, λ_0, $\mathsf{E}X$ and $\mathsf{E}S$ that do not satisfy the negative-drift condition, distributions of the interarrival times, service times, available and unavailable periods can be found such that the system is not stable. However, this does not mean that the system is always unstable if (9) is not satisfied. Hence, this section is concerned with tightening the bounds. This comes at a cost; the negative-drift condition in the following theorem is not in terms of the moments of the various driving random variables, but in terms of the counting function of the service process $N_S(t)$,

$$N_S(t) = \min(k \geq 1 : R_{k+1} > t) \quad t \geq 0\,,$$

with $R_k = S_1 + \cdots + S_k$, $k \geq 1$. In view of the former expression, $N_S(t)$ denotes the number of completed service times in an interval of length t.

Theorem 5. *Assume that assumptions (31), (32) and (36) hold and that the following negative drift condition is satisfied,*

$$\mathsf{E}[N_S(Y)] > \frac{\lambda}{\lambda_0}. \tag{40}$$

Then the queue-size process $\nu = \{\nu(t), t \geq 0\}$ is positive recurrent with respect to the zero-delayed renewal process β.

Proof. By the discrete time scale of the queueing model, the number of departures during an interval is bounded by the length of that interval. In particular, $N_S(Y) \leq Y$ w.p.1 which implies $\mathsf{E}N_S(Y) \leq \mathsf{E}Y < \infty$. Applying the dominated convergence theorem on the sequence $\min(l, N_S(Y))$, $l = 1, 2\ldots$, shows that this sequence converges to $\mathsf{E}[N_S(Y)]$. Hence, in view of the negative drift condition (40), there exist a finite integer K such that,

$$\mathsf{E}[\min(K, N_S(Y))] - \frac{\lambda}{\lambda_0} := \delta_0 > 0\,.$$

We first consider the queue content at the end of cycles. Let $\tilde{\nu}_n = \nu(T_n)$ denote the queue content at the end of the nth cycle and let $\Delta_A(n) = N_A(T_n) - N_A(T_{n-1})$ denote the number of arrivals during this cycle. We have the following recursion,

$$\tilde{\nu}_0 = 1\,, \quad \tilde{\nu}_n = \tilde{\nu}_{n-1} + \Delta_A(n) - N_n\,, \quad n \geq 1, \tag{41}$$

where N_n is the actual number of departures in the interval $(T_{n-1}, T_n]$. For ease of notation, let $S_{n,l}$ denote the lth customer service time of a customer which is served during the nth available period and let

$$B_n = \max\{0 \le k \le K : \sum_{l=1}^{k} S_{n,l} \le Y_n\}\,.$$

One then easily shows that for given A_n, X_n, Y_n and $S_{n,l}$, we have $\tilde{\nu}_n \le u_n$, the latter being defined by the recursion

$$u_{n+1} = (u_n - B_{n+1})^+ + \Delta_A(n+1)\,,$$

for $n \ge 0$ and with $u_0 = 1$. Notice that the numbering of the service times is key to obtain the dominance of u_n. In addition, the sequence $\{u_n\}$ can be interpreted as the queue content at the end of cycles for a queueing system with interruptions, whereby arrivals cannot be served during the cycle in which they arrive and whereby at most K customers are served during a cycle.

By the identity $(x)^+ = x + (-x)^+$, we further find,

$$u_{n+1} = N_A(T_{n+1}) - \sum_{k=0}^{n} B_{k+1} + \sum_{k=0}^{n} (B_{k+1} - u_k)^+\,,$$

such that,

$$\sum_{k=1}^{n} (B_{k+1} - u_k)^+ = u_{n+1} + \sum_{k=0}^{n} B_{k+1} - N_A(T_{n+1}) - (B_1 - 1)^+ \ge \sum_{k=1}^{n} B_{k+1} - N_A(T_{n+1})\,. \tag{42}$$

Let $\mu(n)$ denote the time that the queue size is less than K in the interval $(0, n]$. We obtain the following inequality,

$$\begin{aligned}
\mu(n) &= \sum_{i=1}^{n} 1(\nu(i) < K) \ge \sum_{i=1}^{n} 1(\nu(i) < K, T(i) = 0) \\
&= \sum_{k=1}^{N_Z(n)} 1(\nu(T(k)) < K) \ge \sum_{k=1}^{N_Z(n)} 1(u_k < K) \\
&\ge \sum_{k=1}^{N_Z(n)} 1(u_k < B_{k+1}) \ge \frac{1}{K} \sum_{k=1}^{N_Z(n)} (B_{k+1} - u_k)^+.
\end{aligned}$$

In view of inequality (42), we find the following upper bound for $\mu(n)$,

$$\mu(n) \ge \frac{1}{K} \left(\sum_{k=1}^{N_Z(n)} B_{k+1} - N_A(T_{N_Z(n)+1}) \right). \tag{43}$$

By the SLLN we have w. p. 1,

$$\begin{aligned}
\lim_{n\to\infty} \frac{N_A(T_n)}{T_n} &= \lambda\,, & \lim_{n\to\infty} \frac{T_n}{n} &= \frac{1}{\lambda_0}\,, \\
\lim_{n\to\infty} \frac{N_Z(n)}{n} &= \lambda_0\,, & \lim_{n\to\infty} \frac{\sum_{m=1}^{n} B_m}{n} &= \mathsf{E}[\min(N_S(Y), K)]\,.
\end{aligned} \tag{44}$$

Equation (43) and (44) then yields,

$$\liminf_{n\to\infty} \frac{\mu(n)}{n} \ge \frac{\lambda_0}{K} \mathsf{E}[\min(N_S(Y), K)] - \frac{\lambda}{K} = \frac{\delta_0}{K} > 0.$$

The argument leading to equation (16) of Theorem 1 again applies. Hence, there exist a non-random (sub)sequence of time instants $n_k \to \infty$ (as $k \to \infty$) and some $\varepsilon > 0$ such that

$$\inf_i \mathsf{P}(\nu(n_k) < K) \geq \varepsilon \,.$$

Finally, the arguments of Theorem 4 show the positive recurrence of the queue-size process with respect to β. □

Remark 5. The negative drift condition in the preceding theorem has a simple intuitive interpretation. $\mathsf{E}N_S(Y)$ denotes the mean number of customers that can receive service during an available period. Moreover λ/λ_0 denotes the mean number of customers that arrive during a cycle. Since there is only service during the available periods, $\mathsf{E}N_S(Y)$ also denotes the number of customers that can receive service during a cycle. The negative drift condition thus states that the mean number of possible services during a cycle should exceed the mean number of arrivals during a cycle.

Now we show the relation between the negative drift condition found above and the corresponding condition of Theorem 4. In particular, the following theorem shows that Theorem 5 refines Theorem 4.

Theorem 6. *Assume that (36) and (37) hold. Then (40) is satisfied.*

Proof. In view of the definition of λ_0, the negative drift condition (37) can be rewritten as follows,

$$\frac{\mathsf{E}[Y] - \mathsf{E}[S]}{\mathsf{E}[S]} > \frac{\lambda}{\lambda_0} \,. \tag{45}$$

Wald's equality for renewal processes and the independence of the service and interruption processes further yields,

$$\mathsf{E}[R_{N(Y)+1}] = \mathsf{E}[S]\mathsf{E}[N_S(Y)] + \mathsf{E}[S] \,.$$

Recall that R_i is the ith renewal epoch of the renewal process $N_S(t)$. Further, by definition we have $Y \leq R_{N(Y)+1}$ w.p.1 and therefore also $\mathsf{E}[Y] \leq \mathsf{E}[R_{N(Y)+1}]$. We thus find,

$$\mathsf{E}[N_S(Y)] \geq \frac{\mathsf{E}[Y] - \mathsf{E}[S]}{\mathsf{E}[S]} \,. \tag{46}$$

Combining (45) and (46) yields the stated result. □

Remark 6. Denote the generating function of the service times by $G(z) = \mathsf{E}z^S$ and let $\mathsf{E}Y = 1/\varphi$. Moreover, assume that available periods are geometrically distributed. Then the drift condition (40) simplifies to,

$$\frac{G(1-\varphi)}{(1-\varphi)\,(1-G(1-\varphi))} > \frac{\lambda}{\lambda_0} \,.$$

This condition was already established in [7] as necessary stability condition.

7. Concluding comments

In this paper, we considered stability of queues with preemptive service interruptions. Stability conditions are not trivial since such queueing systems are in general not work-conserving. We first obtained a stability condition based on the workload process. This condition is expressed in terms of the first moments of the arrival, service and interruption processes and therefore easy to evaluate. Also we presented a new approach to extend stability analysis to non-zero initial states. Further, the condition is tight in the sense that, for any given $\lambda_0^{(i)}$, λ, $\mathsf{E}X^{(i)}$ and $\mathsf{E}S$ that do not satisfy the negative-drift condition, distributions of the interarrival times, service times, available and unavailable periods can be found such that the system is not stable. However, this does not mean that the system is always unstable if the condition is not satisfied. We therefore refined the stability condition by focusing on the queue-size process. However, this refinement came at the cost that the stability condition is no longer expressed in terms of the first moments of given variables.

Acknowledgements

The authors are thankful to the referees for careful reading and useful comments which have helped us to improve the readability of the paper.

References

[1] B. Doshi, Queueing systems with vacations – a survey, Queueing Systems 1 (1986) 29–66.
[2] H. White, L. Christie, Queuing with preemptive priorities or with breakdown, Operations Research 6 (1) (1958) 79–95.
[3] B. Avi-Itzhak, P. Naor, Some queuing problems with the service station subject to breakdown, Operations Research 11 (3) (1963) 303–319.
[4] K. Thiruvengadam, Queuing with breakdowns, Operations Research 11 (1) (1963) 62–71.
[5] D. Gaver Jr., A waiting line with interrupted service, including priorities, Journal of the Royal Statistical Society B24 (1962) 73–90.
[6] P. Nain, Queueing systems with service interruptions: an approximate model, Performance Evaluation 3 (2) (1983) 123–129.
[7] D. Fiems, B. Steyaert, H. Bruneel, Discrete-time queues with generally distributed service times and renewal-type server interruptions, Performance Evaluation 55 (3-4) (2004) 277–298.
[8] H. Takagi, A survey of queueing analysis of polling models, in: Proceedings of the Third IFIP International Conference on Data Communication Systems and Their Performance, Rio de Janeiro, Brazil, 1987, pp. 263–281.
[9] E. Altman, D. Fiems, Expected waiting time in polling systems with correlated vacations, Queueing Systems 56 (3–4) (2007) 241–253.
[10] D. Lee, B. Sengupta, An approximate analysis of a cyclic server queue with limited service, Queueing Systems 11 (1992) 153–178.
[11] M. van Vuuren, E. Winands, Iterative approximation of k-limited polling systems, Queueing Systems 55 (2007) 161–178.
[12] K. Leung, D. Lucantoni, Two vacation models for token-ring networks where service is controlled by timers, Performance Evaluation 20 (1–3) (1994) 165–184.
[13] K. Leung, Cyclic-service systems with nonpreemptive, time-limited service, IEEE Transactions on Communications 42 (1994) 2521–2524.
[14] I. Rubin, J. Wu, Analysis of an M/G/1/N queue with vacations and its iterative application to FDDI timed-token rings, IEEE/ACM Transactions on Networking 3 (1995) 842–856.
[15] E. de Souza e Silvia, H. Gail, R. Muntz, Polling systems with sever timeouts and their application to token passing networks, IEEE/ACM Transactions on Networking 3 (1995) 560–575.
[16] I. Frigui, A. Alfa, Analysis of a time-limited polling system, Computer Communications 21 (1998) 558–571.
[17] A. Federgruen, L. Green, Queueing systems with service interruptions, Operations Research 34 (5) (1986) 752–768.
[18] H. Bruneel, A general treatment of discrete-time buffers with one randomly interrupted output line, European Journal of Operational Research 27 (1) (1986) 67–81.
[19] D. Lee, Analysis of a single server queue with semi-Markovian service interruption, Queueing Systems 27 (1–2) (1997) 153–178.
[20] D. Fiems, B. Steyaert, and H. Bruneel, Randomly interrupted GI-G-1 queues: Service strategies and stability issues, Annals of Operations Research, 112(1-4) (2002) 171–183.
[21] D. Fiems, B. Steyaert, and H. Bruneel, Analysis of a discrete-time GI-G-1 queueing model subjected to bursty interruptions, Computers & Operations Research, 30(1) (2003) 139–153.
[22] Y. Tang, A single-server M/G/1 queueing system subject to breakdowns — some reliability and queueing problems, Microelectonics Reliability 37 (1997) 315–321.
[23] W. Li, D. Shi, X. Chao, Reliability analysis of M/G/1 queueing systems with server breakdowns and vacations, Journal of Applied Probability 34 (1997) 546–555.
[24] R. Núñez Queija, Sojourn times in a processor sharing queue with service interruptions, Queueing Systems 34 (1–4) (2000) 351–386.
[25] B. Balcioğlu, D. L. Jagerman, T. Altiok, Approximate mean waiting time in a GI/D/1 queue with autocorrelated times to failures, IIE Transactions 39 (10) (2007) 985–996.
[26] D. Fiems, T. Maertens, H. Bruneel, Queueing systems with different types of interruptions, European Journal of Operations Research 188 (2008) 838–845.
[27] I. Mitrany, B. Avi-Itzhak, A many-server queue with service interruptions, Operations Research 16 (1968) 628–638.
[28] M. Neuts, D. Lucantoni, Markovian queue with N-servers subject to breakdowns and repairs, Management Science 25 (9) (1979) 849–861.
[29] H. Chen, Fluid approximation and stability of multiclass queueing networks: work-conserving disciplines, Annals of Applied Probabability, 5 (1995) 637-665.
[30] H. Chen and D. Yao, Fundamentals of queueing networks, Springer, 2001.
[31] J. Dai, On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models, Annals of Applied Probabability, 5 (1995) 49-77.
[32] S. Foss and T. Konstantopoulos, An overview on some stochastic stability methods, *Journal of the Operations Research Society of Japan*, v. 47, No 4, (2004) 275-303.
[33] E. Morozov, The tightness in the ergodic analysis of regenerative queueing processes, Queueing Systems 27 (1997) 179–203.

[34] E. Morozov, A multiserver retrial queue: regenerative stability analysis, Queueing Systems 56 (2007) 157–168.
[35] S. Asmussen, Applied Probability and Queues, Springer, 2002.
[36] K. Sigman, One-dependent regenerative processes and queues in continuous time, Mathematics of Operations Research 15 (1990) 175–189.
[37] W. Feller, An Introduction to Probability Theory and Its Applications, Vol. II, John Wiley & Sons, 1971.
[38] W. Rogiest, E. Morozov, D. Fiems, H. Bruneel, Stability of single-wavelength optical buffers, European Transactions on Telecommunications, 21 (2010) 202-212.
[39] H. Kaspi, A. Mandelbaum, Regenerative closed queueing networks, Stochastics and Stochastics Reports 39 (1992) 239–258.