Int. J. Appl. Math. Comput. Sci., 2018, Vol. 28, No. 4, 695–704
DOI: 10.2478/amcs-2018-0053

amcs

FROM EXHAUSTIVE VACATION QUEUES TO PREEMPTIVE PRIORITY QUEUES WITH GENERAL INTERARRIVAL TIMES

DIETER FIEMS [a,*], STIJN DE VUYST [b]

[a]Department of Telecommunications and Information Processing
Ghent University, St-Pietersnieuwstraat 41, 9000 Gent, Belgium
e-mail: Dieter.Fiems@UGent.be

[b]Department of Industrial Systems Engineering and Product Design
Ghent University, Technologiepark 3, 9052 Zwijnaarde, Belgium
e-mail: Stijj.DeVuyst@UGent.be

We consider the discrete-time $G/GI/1$ queueing system with multiple exhaustive vacations. By a transform approach, we obtain an expression for the probability generating function of the waiting time of customers in such a system. We then show that the results can be used to assess the performance of $G/GI/1$ queueing systems with server breakdowns as well as that of the low-priority queue of a preemptive $M^X+G/GI/1$ priority queueing system. By calculating service completion times of low-priority customers, various preemptive breakdown/priority disciplines can be studied, including preemptive resume and preemptive repeat, as well as their combinations. We illustrate our approach with some numerical examples.

Keywords: queueing system, preemptive priority, server interruption, server breakdown, exhaustive vacations.

1. Introduction

This paper presents the waiting time analysis of three related discrete-time queueing models. The first model is the $G/GI/1$ exhaustive multiple vacation queueing system. As opposed to ordinary single-server queueing systems, the server of this system does not simply wait for the next arrival when the queue is empty upon departure of a customer. Instead, it leaves for a vacation of random duration. If there is a customer waiting upon returning from this vacation, the server starts working again. If not, the server leaves for another vacation of random duration; see, e.g., the works of Doshi (1986), Takagi (1991) or Tian and Zhang (2006) for references to $M/G/1$-type vacation models. Some recent advances on queues with vacations are presented by Woźniak *et al*. (2014), Dudin *et al*. (2016) and Atencia (2016).

The second model is the $G/GI/1$ queueing system with server interruptions or breakdowns; see the work of Krishnamoorthy *et al*. (2014) for a recent survey on queues with server interruptions. For this queueing system, the server is not continuously available. From time to time, the server breaks down and a random repair time is required to restore proper server behaviour. In particular, it is allowed to break down while a customer receives service. The breakdown discipline then determines how the service goes on after the corresponding repair. The breakdown disciplines studied include repeating and continuing service after the repair, as well as more intricate combinations of these disciplines, which are introduced below.

The third queueing model is a preemptive priority queueing system. In a preemptive priority queue with two classes, say a high- and a low-priority class, customers of the high-priority class are served whenever there are such customers in the system. In particular, when a low-priority customer is being served upon arrival of a high-priority customer, the former immediately leaves the server, making room for the latter. Hence, for preemptive priority systems, the presence of low-priority customers does not affect the performance of high-priority customers. When all high-priority customers have left the system, the server either resumes the service of the low-priority customer (preemptive-resume (see Walraevens *et al*., 2004)), or repeats the service of the customer. In

*Corresponding author

the latter case, the service time may be resampled (preemptive-repeat-different (see Lee and Lee, 2003)), or may remain the same (preemptive-repeat-identical (see Walraevens *et al*., 2006)).

These queueing systems can be described as follows. For all breakdown disciplines, a service-completion-time approach allows reducing that of the queueing system with breakdowns to that of the $G/GI/1$ exhaustive multiple vacation queueing one. Similarly, by identifying the busy periods of high-priority customers with the repair times of the queueing system with breakdowns, the system with breakdowns can be used to assess performance of low-priority customers. Breakdown disciplines then immediately relate to priority disciplines like preemptive resume and preemptive repeat.

1.1. Related work on priority queues. Queueing systems with vacations having been well surveyed (Doshi, 1986; Takagi, 1991; Tian and Zhang, 2006). We discuss related work on preemptive priority queues in this section, and on queues with interruptions in the next one.

Preemptive priority queueing models find applications in diverse fields, including telecommunication networks and production systems. Jayaswal *et al*. (2011) use a preemptive-resume priority queueing system to assess a company's service differentiation strategy in price and delivery times towards two different customer segments in a capacitated environment. Adan *et al*. (2009), study a preemptive-resume queueing model to determine optimal spare-part inventory levels in a repair shop where multiple types of parts are repaired. Whenever a part of a production system is in repair, the system which is costly is offline, the exact cost depending on the part type. The preemptive-resume discipline is also applied to assess performance of cognitive radio networks where secondary users access the wireless channel when it is not employed by its primary users. The priority model allows us to account for channel access by primary users, sensing errors of secondary users, as well as for heterogeneous channel capacity (Wang *et al*., 2011).

For a distributed database system with file-replication, updates of the database files have preemptive-repeat-different priority over database requests in the work of Sumita and Sheng (1988). A preemptive-repeat-identical priority discipline is applied to assess performance of 1- and p_i-persistent CSMA-CD protocols for an unslotted fiber-optic bus network with a finite number of stations (Yoon and Un, 1994). Because of the unidirectional transmission, an upstream station has priority over downstream stations in accessing the channel. In particular, for the 1-persistent CSMA-CD protocol, retransmission of deferred packets starts when the channel is sensed idle, which yields a pure repeat-identical priority discipline. A similar model is studied to assess performance of an optical metropolitan area network by Castel and Hebuterne (2004).

1.2. Related work on queues with interruptions. Interest in queueing models with interruptions dates back to White and Christie (1958), who studied an M/M/1 queueing system with exponential repair times. The generalisation to generally distributed service times and repair times was later performed by Gaver Jr. (1962), Avi-Itzhak and Naor (1963), and Thiruvengadam (1963). Further extensions include phase-type distributions for the available periods (Federgruen and Green, 1986), arrival correlation (Takine and Sengupta, 1997), processor-sharing service (Núñez Queija, 2000), retrials (Dragieva, 2014; Zhang and Zhu, 2013; Gao *et al*., 2016), priorities (Sahba *et al*., 2013) and multi-server systems (Kim *et al*., 2017). Tang (1997) considers Poisson breakdowns when the server is working and renewal type breakdowns when it is idle, whereas Li *et al*. (1997) investigate the transient behaviour of the $M/G/1$ queue subject to Poisson breakdowns. Except for Gaver Jr. (1962), all these authors consider preemptive resume interruptions; service continues after the interruption (we borrow the parlance from priority queueing systems). Gaver Jr. (1962) also considers preemptive repeat and preemptive repeat different service interruptions; in this case the service time is repeated after the interruption with the same service time or with a new resampled service time.

Preemptive repeat and resume are not the only possible interruption strategies. For example, a service can can have multiple phases and only the ongoing phase is repeated after the interruption (with or without identical service time) (Fiems *et al.*, 2002; 2004). Furthermore, the breakdown rate can also depend on the on-going phase (Wang, 2004; Choudhury and Tadj, 2009), or breakdowns can either trigger a preemptive resume or preemptive repeat interruption (Fiems *et al*., 2008). Some breakdown models include additional features such as multiple vacations when the server is idle (Tang *et al*., 2008), setup and closedown times (Ke, 2007), reneging (Martin and Mitrani, 2008), customer expulsions on arrivals (Atencia, 2014; 2015), delayed detection of breakdowns (Krishnamoorthy *et al*., 2015), multiple types of interruptions (Wu *et al*., 2011), or working breakdowns (Jiang and Liu, 2017).

Some authors also consider queues with interruptions and generally distributed interarrival times. Balcioğlu *et al*. (2007) approximate a $GI/D/1$ queue with correlated server breakdowns by the corresponding system with an interruption process with (independent) hyper-exponential on-times and general off-times. Lu *et al*. (2016) study the $G/GI/N$ multiserver queue with interruptions in the Halfin–Whitt regime, that is, when the arrival rate and the number of servers are sent to infinity,

while Pang and Zhou (2016) consider a $G/G/\infty$ queueing model with server interruptions. Finally, sufficient conditions for the stability of a multiserver breakdown model are the subject of the work of Morozov *et al.* (2011).

1.3. Organisation of the paper. The remainder of this contribution is organised as follows. In the following section, we describe the $G/GI/1$ queueing system with exhaustive vacations and derive an expression for the probability generating function of the waiting times. Section 3 is then concerned with the interruption model. The model details are introduced and expressions for the probability generating function of the service completion times are obtained. The probability generating function of the waiting times of the interruption model is then obtained by combining the expression of the service completion times with the waiting time analysis of the exhaustive vacation queue. The third model at hand, the preemptive priority queue, is the subject of Section 4. A high-priority busy period analysis transforms the low-priority queue into a queue with interruptions; the results of Section 3 can thus be applied to study waiting times of low-priority customers. Finally, we draw conclusions in Section 6.

2. G/GI/1 queue with exhaustive vacations

2.1. Queueing model and assumptions. We consider a discrete-time queueing system, i.e., we assume that time is divided into fixed length intervals or slots. At the consecutive slot boundaries, customers arrive at the system, are stored in an infinite capacity queue, and are served on a first-come-first-served basis. No more than one customer can be served at the same time, and service times are integer multiples of the slot length such that all departures occur at slot boundaries as well.

There is at most one customer arrival at a slot boundary. The inter-arrival times between consecutive customers (expressed in numbers of slots) constitute a sequence of independent and identically distributed (i.i.d.) positive random variables with a common probability generating function $A(z)$ which is assumed to be a rational function. As $A(z)$ is rational, there exists a real-valued $a^* > 1$ such that $A(z)$ is analytic in $|z| < a^*$. For further use, let a be a constant such that $1 < a < a^*$. In addition, the service times of the consecutive customers constitute a sequence of i.i.d. positive random variables as well; let $S(z)$ denote the common probability generating function of the service times.

As long as there are other customers in the queue, the server serves one customer after another. However, if the queue is empty upon departure of a customer, the server leaves for a vacation of random duration. If there are customers in the queue when the server returns from its vacation, the server immediately resumes serving customers. If this is not the case, it immediately leaves for another vacation. That is, the server takes consecutive vacations and service is only resumed when the server finds waiting customers upon returning from a vacation. The consecutive vacation times constitute a sequence of i.i.d. positive random variables. Let $V(z)$ denote their common probability generating function.

Finally, let $\bar{A} = A'(1)$, $\bar{S} = S'(1)$ and $\bar{V} = V'(1)$ denote the mean interarrival time, the mean service time and the mean vacation time, respectively.

2.2. Customer waiting time analysis. Let customer waiting time be defined as the number of slots between customer arrival and the slot boundary where the customer enters the server, and let W_k denote the waiting time of the k-th customer. In accordance with the exhaustive multiple vacation system, consecutive customer waiting times are described as

$$W_{k+1} = \begin{cases} W_k + X_k & \text{if } W_k + X_k \geq 0, \\ V_k^{[R]}(-W_k - X_k) & \text{if } W_k + X_k < 0, \end{cases} \tag{1}$$

with $X_k \doteq S_k - A_k$. Here A_k denotes the inter-arrival time between customer k and customer $k+1$, and S_k denotes the service time of customer k. Further, $V_k^{[R]}(i)$ is a random variable that denotes the remaining vacation time upon arrival of a customer that arrives at the i-th slot boundary since the queue became empty. The first equation also holds for the $GI/G/1$ queue without vacations and corresponds to the case where customer $k+1$ arrives in a non-empty queue. The second equation holds if the queue is empty upon departure of customer k. The waiting time of customer $k+1$ then consists of the remaining vacation time upon arrival of this customer.

Let $W_k(z)$ denote the probability generating function of the waiting time of customer k. Moreover, we introduce the annulus $\mathcal{N} = \{z \in \mathbb{C}; a^{-1} < |z| < 1\}$. In view of (1) and by means of some standard z-transform manipulations, we find

$$W_{k+1}(z) = W_k(z)A(1/z)S(z) + \mathrm{E}[(z^{V_k^{[R]}(U_k)+U_k} - 1)z^{-U_k}1\{U_k > 0\}], \tag{2}$$

for $z \in \mathcal{N}$. Here we introduced the random variable $U_k \doteq A_k - W_k - S_k$ for ease of notation.

Let $\theta_i(n)$, $(n = 1, 2, \ldots)$ and $\Theta_i(z)$ denote the probability mass function and the probability generating function of $V_k^{[R]}(i)$, respectively. By conditioning on the length of the first vacation, we find

$$\theta_i(n) = v(i+n) + \sum_{j=1}^{i-1} v(j)\theta_{i-j}(n), \tag{3}$$

with $v(n)$ ($n > 0$) being the probability mass function of the vacation times. The corresponding probability generating function then satisfies

$$\Theta_i(z)z^i - 1 = V(z) - 1 + \sum_{j=1}^{i-1} v(j)\left(\Theta_{i-j}(z)z^{i-j} - 1\right) z^j , \quad (4)$$

for $z \in \mathcal{D} = \{z \in \mathbb{C}; |z| < 1\}$. Noting that $V(z) - 1$ has no zeroes inside the unit disk, introduce the functions

$$\Omega_i(z) = \frac{\Theta_i(z)z^i - 1}{V(z) - 1} ,$$

for $z \in \mathcal{D}$. Solving the former expression for $\Theta_i(z)$ and substituting the result in (4) yield,

$$\Omega_i(z) = 1 + \sum_{j=1}^{i-1} v(j)\Omega_{i-j}(z)z^j ,$$

for $z \in \mathcal{D}$. We have $\Omega_1(z) = 1$, $\Omega_2(z) = 1 + v(1)z$, etc. In view of the former expression, one easily verifies by recursion that $\Omega_i(z)$ is a polynomial of order $i - 1$ coefficients in $[0, 1]$ which depend on the probability mass function of the vacations. Moreover, the coefficients of the terms in the polynomial are non-negative and bounded by 1.

As $\Theta_i(z)$ is the probability generating function of $V^{[R]}(i)$, we have, for $z \in \mathcal{D}$,

$$\begin{aligned} \mathrm{E}[z^{V^{[R]}(i)+i} - 1] &= \Theta_i(z)z^i - 1 \\ &= (V(z) - 1)\Omega_i(z) . \end{aligned}$$

Combining this expression with (2) then yields

$$W_{k+1}(z) \quad (5)$$
$$= W_k(z)A(1/z)S(z) \quad (6)$$
$$+ (V(z) - 1)z^{-1}\,\mathrm{E}[\Omega_{U_k}(z)z^{-U_k+1}1\{U_k > 0\}] , \quad (7)$$

for $z \in \mathcal{N}$.

Under the assumption that the queueing system under consideration reaches a steady state, a standard Loynes argument shows that this is the case if the arrival load does not exceed the service capacity, i.e., for $\bar{A} > \bar{S}$ — let $W(z)$ denote the steady state probability generating function of the customer waiting time. We find that

$$W(z)(1 - A(1/z)S(z))z = (V(z) - 1)\Upsilon(1/z) , \quad (8)$$

for $z \in \mathcal{N}$. Here $\Upsilon(z)$ is a z-transform with positive coefficients υ_n,

$$\Upsilon(z) = \sum_{n=0}^{\infty} \upsilon_n z^n = \mathrm{E}[\Omega_U(1/z)z^{U-1}1\{U > 0\}] ,$$

where U denotes the steady state version of U_k.

Clearly, $A(1/z)$ is rational since $A(z)$ is rational. Therefore, let $P_A(z)$ and $Q_A(z)$ denote respectively the numerator and the denominator of $A(1/z)$,

$$A(1/z) = \frac{P_A(z)}{Q_A(z)} . \quad (9)$$

The numerator and the denominator are uniquely defined up to a factor. Moreover, we have $P_A(1) = Q_A(1)$ by the normalisation condition $A(1) = 1$. In the remainder, we choose $P_A(1) = Q_A(1) = 1$ to simplify notation (but any non-zero constant yields the same results).

Combining (8) and (9) further yields

$$W(z)\frac{Q_A(z) - P_A(z)S(z)}{V(z) - 1} = \frac{Q_A(z)}{z}\Upsilon(1/z) , \quad (10)$$

for $z \in \mathcal{N}$.

We now study the analyticity of both sides in (10). The left-hand side of the former equation is analytic within the unit disk $\mathcal{D}$. Indeed, $Q_A(z)$ and $P_A(z)$ are entire functions, $V(z) - 1$ has no zeroes in $\mathcal{D}$, while the probability generating functions $W(z)$, $S(z)$ and $V(z)$ are analytic in $\mathcal{D}$.

For the right-hand side, we need to study $\Upsilon(z)$ in detail. As $\Omega_i(z)$ is a polynomial of order $i - 1$ with coefficients in $[0, 1]$, so is $\Omega_i(1/z)z^{i-1}$. Therefore, we have $|\Omega_i(1/z)z^{i-1}| \leq ia^{i-1}$ for $|z| < a$. This in turn implies

$$\begin{aligned} |\Upsilon(z)| &\leq \mathrm{E}[|\Omega_U(1/z)z^{U-1}|1\{U > 0\}] \\ &\leq \mathrm{E}[Ua^{U-1}1\{U > 0\}] \\ &\leq \mathrm{E}[Aa^{A-1}1\{A > W + S\}] \\ &\leq \mathrm{E}[Aa^{A-1}] = A'(a) < \infty . \end{aligned}$$

The first inequality follows from Jensen's inequality, the second from the bound on $|\Omega_i(1/z)z^{i-1}|$ found above, and the third from $U = A - W - S \leq A$. As a consequence, we find that the power series $\sum \upsilon_n z^n$ converges for $|z| < a$, which implies that $\Upsilon(z)$ is analytic in this region. Summarising, $\Upsilon(1/z)$ is analytic for $|z| > a^{-1}$, and so is the right-hand side of (10) as $Q_A(z)$ is entire and z^{-1} is analytic for $|z| > 0$.

Thus the left and right-hand sides are analytic functions in $|z| < 1$ and $|z| > a^{-1}$, respectively and equal for $a^{-1} < |z| < 1$. As such, the left and right-hand sides are analytic continuations of each other. That is, the left and right-hand sides are representations of an entire function $\Phi(z)$ in their respective domains.

The left-hand side of (10) is bounded for $|z| \leq a^{-1}$. Moreover, $\Upsilon(1/z)$ is analytic and bounded for $|z| \geq a^{-1}$. $Q_A(z)$ being a polynomial of degree q_A, we find that both the left- and the right-hand side of (10) satisfy the inequality

$$|\Phi(z)| \leq \alpha + \beta|z|^{q_A - 1} , \quad (11)$$

for some constants α and β and for all $z \in \mathbb{C}$.

In accordance with the extended Liouville theorem (see, e.g., Theorem 5.11 of Bak and Newman (1997)), this inequality implies that $\Phi(z)$ is a polynomial of degree at most $q_A - 1$. Summarising, we have, for all z inside the unit disk,

$$W(z) = \frac{(V(z)-1)\Phi(z)}{Q_A(z) - P_A(z)S(z)}, \tag{12}$$

where $\Phi(z)$ is an unknown polynomial of degree at most $q_A - 1$.

By Klimenok's theorem (Klimenok, 2001), one shows that the denominator of (12) has $q_A - 1$ zeros, say $z = y_1, \ldots, y_{q_A-1}$, inside the unit disk. Since probability generating functions are analytic inside the unit disk, every zero of the denominator is also a zero of the numerator. This then implies

$$\Phi(z) = K \prod_{j=1}^{q_A-1} \frac{z - y_j}{1 - y_j}, \tag{13}$$

where K is an unknown constant. Finally, plugging (13) in (12) and invoking the normalisation condition yield,

$$W(z) = \frac{\bar{A} - \bar{S}}{\bar{V}} \frac{V(z) - 1}{Q_A(z) - P_A(z)S(z)} \prod_{j=1}^{q_A-1} \frac{z - y_j}{1 - y_j}. \tag{14}$$

By the moment generating property of the probability generating function, one easily obtains expressions for the various moments of the customer waiting time. In particular, the mean waiting time $\bar{W} = W'(1)$ equals

$$\bar{W} = \sum_{j=1}^{q_A-1} \frac{1}{1 - y_j} + \frac{A''(1) + 2\bar{A}(1 - \bar{A}) + S''(1)}{2(\bar{A} - \bar{S})} + \frac{V''(1)}{2\bar{V}} - P_A'(1). \tag{15}$$

3. G/GI/1 queue with server interruptions

3.1. Modelling assumptions. We retain the assumptions of Section 2 on the discrete-time setting as well as on the arrival process. As opposed to Section 2, the server now does not leave for vacations when the queue becomes empty. In addition, the assumptions on the service process are refined as follows.

Customer service consists of K consecutive stages, let $S_{k,i}$ denote the length of the i-th stage of the service time of customer k such that the vector $S_k \doteq [S_{k,i}]_{i=1,\ldots,K}$ characterises the service requirement of the k-th customer. The consecutive S_k constitute a sequence of independent and identically distributed non-negative random vectors with common probability mass function $s(n_1, n_2, \ldots, n_K)$. For further use, let $S_i(z)$ denote the probability generating function of the i-th part,

$$S_i(z) = \mathrm{E}[z^{S_{k,i}}].$$

There is a single server which can break down from time to time. In particular, we assume that the server breaks down at the end of a slot with a fixed probability α or remains available with probability $\bar{\alpha} \doteq 1 - \alpha$, independent of the state of the queue. After a breakdown, the server needs to be repaired. The consecutive repair times constitute a sequence of independent and identically distributed random variables with a common probability generating function $B(z)$ and a common mean value $\bar{B} \doteq B'(1)$. There are no additional breakdowns during repair. In other words, the server alternates between repair and availability, availability periods being (shifted) geometrically distributed.

Clearly, the server can break down while a customer is being served. In this case, service of the stage is either resumed or repeated after the repair, depending on the sequence number of the stage. If service is repeated, the required service time remains the same over the consecutive trials; this is a preemptive repeat identical service discipline. Without loss of generality, we assume that service of the first stage resumes after a breakdown, while service is repeated after a breakdown during the other stages. This indeed does not compromise generality: the order of the consecutive stages does not affect the performance measures under study. Moreover, having multiple consecutive stages with preemptive resume is equivalent to having a single stage where the service time of the single stage equals the sum of the service times of the consecutive stages.

Finally, note that the breakdown policy defined above is a generalisation of both preemptive resume and the preemptive repeat identical breakdown policies (Fiems *et al.*, 2004).

3.2. Completion times. To simplify the queueing analysis, we first consider the completion time of a customer. It is defined as the time (in slots) that it actually takes to serve a customer, including repair times and (possible) lost service time. The completion time of a customer starts at the beginning of the slot where the customer receives service for the first time and ends at the beginning of the slot where the server is available to serve the next customer (if present). Hence, a customer's completion time ends at the departure epoch of this customer if the server is available during the slot following the departure, or ends after the repair which starts at the departure epoch if this is not the case.

Let $C_1(z; n)$ denote the probability generating function of the completion time of a customer whose service time equals n slots, assuming a preemptive resume breakdown policy. Clearly, in preemptive resume, every slot of service is followed by a repair time with probability α or by the next service slot with probability $\bar{\alpha}$. Assuming

n slots of service, we immediately find that

$$C_1(z;n) = z^n(\bar{\alpha} + \alpha B(z))^n$$

and

$$C_1'(1;n) = n(1 + \alpha\bar{B})\,.$$

Let $C_2(z;n)$ denote the probability generating function of the completion time of a customer whose service time equals n slots, assuming a preemptive repeat (identical) breakdown policy. By conditioning on the number of slots till the first breakdown, we find

$$C_2(z;n) = \sum_{k=1}^{n-1} \bar{\alpha}^{k-1}\alpha z^k B(z) C_2(z;n) + \bar{\alpha}^{n-1} z^n(\bar{\alpha} + \alpha B(z))\,.$$

Here we used the fact that, by the lack of memory of the geometric time till the next breakdown, consecutive trials are independent. Moreover, as service has to start all over, the remaining completion time after the breakdown is distributed as the service completion time. Solving for $C_2(z;n)$ further yields

$$C_2(z;n) = \frac{\bar{\alpha}^{n-1} z^n (1 - \bar{\alpha}z)(\bar{\alpha} + \alpha B(z))}{1 - \bar{\alpha}z - \alpha z\, B(z)(1 - \bar{\alpha}^{n-1} z^{n-1})}$$

and

$$C_2'(1;n) = \frac{(1 - \bar{\alpha}^n)\,(\alpha\, B'\,(1) + 1)}{\alpha\,\bar{\alpha}^{n-1}}\,.$$

Given $n_1, \ldots, n_K$, the completion times of the different parts are independent. Hence, the probability generating function of the completion time equals

$$C(z) = \sum_{n_1,\ldots,n_K} s(n_1, n_2, \ldots, n_K) C_1(z;n_1) \prod_{k=2}^{K} C_2(z;n_k)\,. \tag{16}$$

By the moment generating property of probability generating functions, the mean service completion time $\bar{C} \doteq C'(1)$ is given by

$$\bar{C} = (1 + \alpha\bar{B}) \left(S_1'(1) + \frac{\bar{\alpha}}{\alpha} \sum_{k=2}^{K} (S_k(\bar{\alpha}^{-1}) - 1) \right)\,.$$

3.3. Waiting time analysis. Having established the expressions of the generating function and moments of the completion times, we now describe the model with server breakdowns to the exhaustive vacation queue. As before, let W_k and A_k denote the waiting time of customer k and the interarrival time between customer k and $k+1$, respectively. Denote by C_k the completion time of the k-th customer. Then

$$W_{k+1} = \begin{cases} W_k + Y_k & \text{if } W_k + Y_k \geq 0, \\ \tilde{V}_k^{[R]}(-W_k - Y_k) & \text{if } W_k + Y_k < 0, \end{cases} \tag{17}$$

with $Y_k \doteq C_k - A_k$, where $\tilde{V}_k^{[R]}(x)$ denotes the remaining repair time at the x-th slot boundary after the departure of the k-th customer. From Section 2, recall that the server does not resume service for a random number of slots when it leaves for a server vacation. Breakdowns while the server is idle affect the server in a similar way. Whenever there is a breakdown in a slot, the server only returns after the corresponding repair or returns the next slot when this is not the case. Hence, we can account for the breakdown and repair process by assuming vacations of length 1 with probability $1-\alpha$ and of length $B+1$ with probability α. Here, B denotes a generic repair time. The probability generating function of these vacations then equals

$$\tilde{V}(z) = (1 - \alpha)z + \alpha z B(z)\,.$$

Summarising, replacing the service times by the completion times and assuming vacations as defined above, we immediately find that the probability generating function of the waiting times and the the mean waiting time in the queueing system with breakdowns are given by

$$W(z) = \frac{\bar{A} - \bar{C}}{1 + \alpha\bar{B}} \frac{z - 1 + \alpha z(B(z) - 1)}{Q_A(z) - P_A(z)C(z)} \prod_{j=1}^{q_A - 1} \frac{z - y_j}{1 - y_j} \tag{18}$$

and

$$\bar{W} = \sum_{j=1}^{q_A - 1} \frac{1}{1 - y_j} + \frac{A''(1) + 2\bar{A}(1 - \bar{A}) + C''(1)}{2(\bar{A} - C'(1))} + \frac{2\alpha\bar{B} + \alpha B''(1)}{2 + 2\alpha\bar{B}} - P_A'(1)\,. \tag{19}$$

Here, the unknowns y_j are the solutions of $Q_A(z) - P_A(z)C(z) = 0$ within the unit disk.

4. Preemptive priority queue

The breakdown model can be used to study waiting times of low-priority customers in a preemptive priority queueing system. By preemption, the server is unavailable whenever there are high-priority customers in the system and available otherwise. In other words, the busy periods of the high-priority queue correspond to the repair periods, and the idle periods of the high-priority queue correspond to the available periods seen by low-priority customers. In view of the preceding sections, we have the following assumptions on low priority customers:

- the sequence of interarrival times between consecutive low-priority customers constitutes a sequence of iid random variables with common probability generating function $A(z)$;

- the refined model for the multi-stage service times of the preceding section holds, and $s(n_1, n_2, \ldots, n_K)$ is the common probability mass function of the stages of the low-priority service times;
- service of the stage is either resumed or repeated after preemption by high priority customers, as in the preceding section.

Recall from the preceding section that the repair periods and available periods constitute sequences of independent generally distributed and geometrically distributed random variables, respectively. Hence, the breakdown model can assess waiting times of low-priority customers if its busy and idle periods fit the modelling assumptions of the breakdown process. This is the case if the high-priority arrival process regenerates whenever there are no arrivals in a slot. Obviously, this is the case when the number of arrivals in consecutive slots and their service times constitute sequences of i.i.d. random variables. Let $A_1(z)$ be the common probability generating function of the number of high-priority arrivals in a slot, and let $S_1(z)$ be the common probability generating function of the service times. The probability generating function $B(z)$ of the busy periods satisfies the following functional equation (Fiems *et al*., 2004):

$$B(z) = \frac{A_1(B_s(z)) - A_1(0)}{1 - A_1(0)}\,, \tag{20}$$

$$B_s(z) = S_1(zA_1(B_s(z))\,, \tag{21}$$

whereas the high-priority queue does not remain empty with probability

$$\alpha = 1 - A_1(0)\,. \tag{22}$$

Remark 1. (*Independence of the arrival process*) The independence assumption of the high-priority arrival process is not required. Regeneration when there are no arrivals is the essential property needed. For example, $M/G/\infty$-input and discrete autoregressive arrival processes of order 1 also regenerate when there are no arrivals.

Summarising, to model the M^X+ $G/G/1$ priority queue, (20), (21) and (22) determine the parameters of the interruption process in Section 3. In other words, plugging (21) and (22) in (18) yields the probability generating function of the waiting times of low-priority customers in the M^X+$G/G/1$ priority queue with the above mentioned preemptive priority policy.

Remark 2. (*Numerical evaluation*) For the evaluation of the moments of the low-priority waiting times, the main numerical difficulty is related to finding the roots y_j in (18). The moments can be expressed in terms of the parameters of the model and these unknown roots; see, e.g., (19) for the mean waiting time, noting that the derivatives of $C(z)$ for $z = 1$ can be obtained in terms of the model parameters. These roots are the zeros of $Q_A(z) - P_A(z)C(z) = 0$, where $C(z)$ is expressed as an infinite sum; see (16). Hence, as we need to evaluate $C(z)$ for $z \neq 1$ for finding the roots, we need to truncate the sum. Moreover, to evaluate the terms in $C(z)$, we need to evaluate $B(z)$ as well, which is implicitly defined in (20) and (21). To find $B(z)$ for $|z| < 1$, we find the root x $(= B_s(z))$ with $|x| < 1$ that solves $x = S_1(zA_1(x))$.

5. Numerical example

To illustrate our results, we now study the influence of the interarrival time distribution of the low-priority traffic on the low-priority waiting time. We assume that the low-priority service time has two stages, and apply preemptive resume and preemptive repeat in the first and second stages, respectively. The length of the first and the second stage are independent and geometrically distributed with parameters β_1 and β_2, respectively, such that the generating function of the completion times simplifies to

$$C(z) = \frac{(1-\beta_1)(1-\beta_2)}{1-\beta_1 z(\bar{\alpha} + \alpha B(z))} \times \sum_{n=1}^{\infty} \beta_2^{n-1} \frac{\bar{\alpha}^{n-1} z^{n+1} (1-\bar{\alpha}z)(\bar{\alpha} + \alpha B(z))^2}{1 - \bar{\alpha}z - \alpha z\, B(z)(1 - \bar{\alpha}^{n-1} z^{n-1})}\,.$$

The batch size of the high-priority (class 1) customers is assumed to be Poisson with rate λ_1, and the high-priority service times are geometrically distributed with mean $\bar{S}_1$,

$$A_1(z) = \exp(\lambda_1(z-1))$$
$$S_1(z) = \frac{z}{\bar{S}_1(1-z) + z}\,.$$

Hence we have $\alpha = 1 - \exp(-\lambda_1)$ and the following functional equation for $B_s(z)$:

$$B_s(z)(\bar{S}_1 \exp(\lambda_1(1 - B_s(z))) - (\bar{S}_1 - 1)z) - z = 0\,.$$

For the low-priority interarrival times, we rely on the moment characterisation of acyclic Coxian phase-type distributions by Horváth *et al*. (2015). That is, we consider the class of distributions with generating function

$$A(z) = p\frac{\phi_1 z}{1-(1-\phi_1)z}\frac{\phi_2 z}{1-(1-\phi_2)z} + (1-p)\frac{\phi_2 z}{1-(1-\phi_2)z}\,,$$

where p, ϕ_1 and ϕ_2 are determined by the first three moments of the interarrival time distribution. Given the mean interarrival time $\mathrm{E}[A]$ and the variance of the

interarrival time $\mathrm{var}[A] \geq \mathrm{E}[A]^2/2$, we find a unique set of parameters (p, ϕ_1, ϕ_2) of the distribution by choosing the distribution with minimal skewness. Hence, the interarrival times are completely characterised by the low-priority (class 2) arrival rate $\lambda_2 = 1/\mathrm{E}[A]$ and the standard deviation $\sigma_2 = \sqrt{\mathrm{var}[A]}$ of the class 2 interarrival times.

Figure 1 investigates the effect of the class 1 arrival rate on the mean class 2 waiting time. The mean class 1 service time equals $\bar{S}_1 = 5$ slots, while both phases of the class 2 service time take 4 slots on the average ($\beta_1 = \beta_2 = 0.75$). The mean class 2 arrival rate is $\lambda_2 = 0.05$, and different values for the standard deviation of the class 2 interarrival times are assumed as depicted. As expected, an increase in the class 1 arrival rate results in an increase in the mean class 2 waiting time. Indeed if the class 1 arrival rate increases, the server needs to attend the class 1 customers more regularly, and the class 2 customers have to wait longer. In addition, the figure readily shows that more variance in the arrival process (increasing values of σ_2) affects the performance negatively.

The latter observation is also confirmed by Fig. 2, which depicts the mean class 2 waiting time $\bar{W}$ vs. the standard deviation σ_2 of the class 2 interarrival times. As above, both phases of the class 2 service time take 4 slots on average ($\beta_1 = \beta_2 = 0.75$). We consider various values of $\bar{S}_1$ as indicated, and set $\lambda_1 = 0.25/\bar{S}_1$ such that the class 1 load is equal for all plots. As expected, the mean class 2 waiting time increases when the standard deviation of the class 2 interarrival times increases.

The effect of the class 1 service times on the class 2 performance is, however, not straightforward. An increase in the mean class 1 service time may either lead to an increase or to a decrease of the mean class 2 waiting time. This observation is clarified in Fig. 3 which depicts the mean class 2 waiting time vs. the mean class 1 service time $\bar{S}_1$, for different values of σ_2 as depicted. The class 1 load is again constant, we set $\lambda_1 = 0.25/\bar{S}_1$ and further retain the parameters of Fig. 3. The figure shows that the mean waiting time first decreases and then increases again for increasing $\bar{S}_1$. This can be explained by noting that, for small $\bar{S}_1$, class 2 service is often interrupted, the server being available only for short times. Indeed, small $\bar{S}_1$ implies large λ_1 and large α, which means that the server often alternates between being available and unavailable for serving class 2 customers. As service interruptions lead to service repetitions, the mean waiting time is large. Performance then improves if one increases $\bar{S}_1$. For high $\bar{S}_1$, we have long periods when the server is available followed by long periods that the server is not available. While there are fewer service interruptions and therefore also fewer service repetitions, the time when the server is not available increases. During this time the class 2 queue builds up, which results in longer class 2 waiting times.

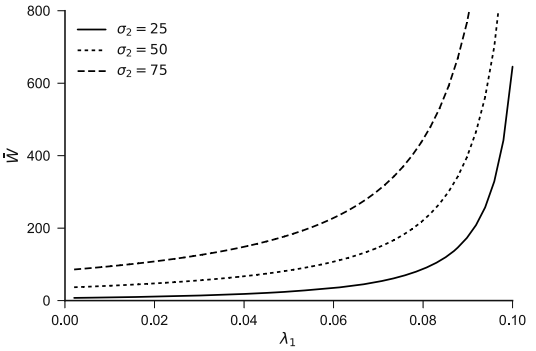

Fig. 1. Mean class 2 waiting time $\bar{W}$ vs. the class 1 arrival rate λ_1 for various values of the standard deviation of the class 2 interarrival times.

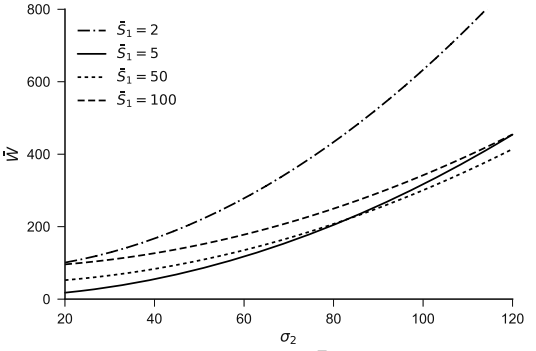

Fig. 2. Mean class 2 waiting time $\bar{W}$ vs. the standard deviation σ_2 of the class 2 interarrival times for different values of the mean class 1 service time $\bar{S}_1$.

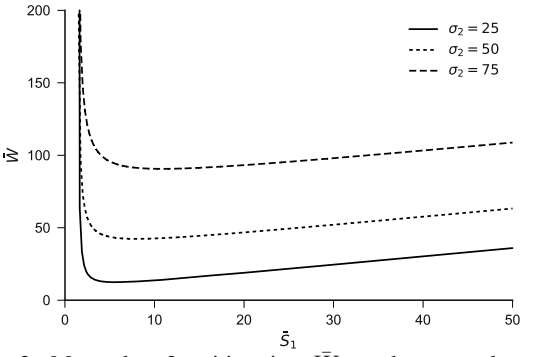

Fig. 3. Mean class 2 waiting time $\bar{W}$ vs. the mean class 1 service time $\bar{S}_1$ for different values of the standard deviation σ_2 of the class 2 interarrival times.

6. Conclusions

This paper considered three related queueing models. Starting off with the $G/GI/1$ queue with exhaustive vacations, it was shown that a queueing system with server interruptions reduces to such a vacation queue, by combining a service completion time approach with particular assumptions on the vacation times. In turn, the queue with service interruptions can be used to

study the performance of low-priority customers in a preemptive priority queueing system. While the relation between interruption queues and priority queues has been established before (Fiems *et al*., 2004), that between vacation queues and interruption queues is new. Moreover, while vacation queues have been studied intensively, we are not aware of a study of the multiple vacation queue in a $G/GI/1$ setting. In fact, as the analysis shows, it is not trivial that arguments for $G/GI/1$ queueing systems extend to equivalent queues with multiple vacations. A slight modification like a first exceptional vacation time already voids the current argument.

References

Adan, I.J.B.F., Sleptchenko, A. and Van Houtum, G.J. (2009). Reducing costs of spare parts supply systems via static priorities, *Asia-Pacific Journal of Operational Research* **26**(4): 559–585.

Atencia, I. (2014). A discrete-time system with service control and repairs, *International Journal of Applied Mathematics and Computer Science* **24**(3): 471–484, DOI: 10.2478/amcs-2014-0035.

Atencia, I. (2015). A discrete-time queueing system with server breakdowns and changes in the repair times, *Annals of Operations Research* **235**(1): 37–49.

Atencia, I. (2016). A discrete-time queueing system with changes in the vacation times, *International Journal of Applied Mathematics and Computer Science* **26**(2): 379–390, DOI: 10.1515/amcs-2016-0027.

Avi-Itzhak, B. and Naor, P. (1963). Some queuing problems with the service station subject to breakdown, *Operations Research* **11**(3): 303–319.

Bak, J. and Newman, D. (1997). *Complex Analysis*, 2nd Edn., Springer-Verlag, New York, NY.

Balcioğlu, B., Jagerman, D.L. and Altiok, T. (2007). Approximate mean waiting time in a $GI/D/1$ queue with autocorrelated times to failures, *IIE Transactions* **39**(10): 985–996.

Castel, H. and Hebuterne, G. (2004). Performance analysis of an optical MAN ring for asynchronous variable length packets, *in* J.N. de Souza *et al.* (Eds.), *Telecommunications and Networking—ICT 2004*, Lecture Notes in Computer Science, Vol. 3124, Springer Verlag, Berlin/Heidelberg, pp. 214–220.

Choudhury, G. and Tadj, L. (2009). An M/G/1 queue with two phases of service subject to the server breakdown and delayed repair, *Applied Mathematical Modelling* **33**(6): 2699–2709.

Doshi, B. (1986). Queueing systems with vacations—a survey, *Queueing Systems* **1**(1): 29–66.

Dragieva, V. (2014). Number of retrials in a finite source retrial queue with unreliable server, *Asia-Pacific Journal of Operational Research* **31**(2), Paper no.: 1440005.

Dudin, A., Moon, H.L. and Dudin, S. (2016). Optimization of the service strategy in a queueing system with energy harvesting and customers' impatience, *International Journal of Applied Mathematics and Computer Science* **26**(2): 367–378, DOI: 10.1515/amcs-2016-0026.

Federgruen, A. and Green, L. (1986). Queueing systems with service interruptions, *Operations Research* **34**(5): 752–768.

Fiems, D., Maertens, T. and Bruneel, H. (2008). Queueing systems with different types of interruptions, *European Journal of Operations Research* **188**(3): 838–845.

Fiems, D., Steyaert, B. and Bruneel, H. (2002). Randomly interrupted $GI - G - 1$ queues: Service strategies and stability issues, *Annals of Operations Research* **112**(1–4): 171–183.

Fiems, D., Steyaert, B. and Bruneel, H. (2004). Discrete-time queues with generally distributed service times and renewal-type server interruptions, *Performance Evaluation* **55**(3–4): 277–298.

Gao, S., Wang, J.T. and Van Do, T. (2016). A repairable retrial queue under Bernoulli schedule and general retrial policy, *Annals of Operations Research* **247**(1): 169–192.

Gaver Jr., D. (1962). A waiting line with interrupted service, including priorities, *Journal of the Royal Statistical Society* **B24**: 73–90.

Horváth, I., Papp, J. and Telek, M. (2015). On the canonical representation of order 3 discrete phase type distributions, *Electronic Notes in Theoretical Computer Science* **318**: 143–158.

Jayaswal, S., Jewkes, E. and Ray, S. (2011). Product differentiation and operations strategy in a capacitated environment, *European Journal of Operational Research* **210**(3): 716–728.

Jiang, T. and Liu, L. (2017). The $GI/M/1$ queue in a multi-phase service environment with disasters and working breakdowns, *International Journal of Computer Mathematics* **94**(4): 707–726.

Ke, J. (2007). Batch arrival queues under vacation policies with server breakdowns and startup/closedown times, *Applied Mathematical Modelling* **31**(7): 1282–1292.

Kim, C., Klimenok, V.I. and Dudin, A.N. (2017). Analysis of unreliable BMAP/PH/N type queue with Markovian flow of breakdowns, *Applied Mathematics and Computation* **314**: 154–172.

Klimenok, V. (2001). On the modification of Rouché's theorem for the queueing theory problems, *Queueing Systems* **38**(4): 431–434.

Krishnamoorthy, A., Jaya, S. and Lakshmy, B. (2015). Queues with interruption in random environment, *Annals of Operations Research* **233**(1): 201–219.

Krishnamoorthy, A., Pramod, P.K. and Chakravarthy, S.R. (2014). Queues with interruptions: A survey, *TOP* **22**(1): 290–320.

Lee, Y. and Lee, K. (2003). Discrete-time $Geo^X/G/1$ queue with preemptive repeat different priority, *Queueing Systems* **44**(4): 399–411.

Li, W., Shi, D. and Chao, X. (1997). Reliability analysis of $M/G/1$ queueing systems with server breakdowns and vacations, *Journal of Applied Probability* **34**(2): 546–555.

Lu, H., Pang, G. and Zhou, Y. (2016). $G/GI/N$ (plus GI) queues with service interruptions in the Halfin–Whitt regime, *Mathematical Methods of Operations Research* **83**(1): 127–160.

Martin, S. and Mitrani, I. (2008). Analysis of job transfer policies in systems withunreliable servers, *Annals of Operations Research* **162**(1): 127–141.

Morozov, E., Fiems, D. and Bruneel, H. (2011). Stability analysis of multiserver discrete-time queueing systems with renewal type server interruptions, *Performance Evaluation* **68**(12): 1261–1275.

Núñez Queija, R. (2000). Sojourn times in a processor sharing queue with service interruptions, *Queueing Systems* **34**(1–4): 351–386.

Pang, G. and Zhou, Y. (2016). $G/G/\infty$ queues with renewal alternating interruptions, *Advances in Applied Probability* **48**(3): 812–831.

Sahba, P., Balcioğlu, B. and Banjevic, D. (2013). Analysis of the finite-source multiclass priority queue with an unreliable server and setup time, *Naval Research Logistics* **60**(4): 331–342.

Sumita, U. and Sheng, O. (1988). Analysis of query processing in distributed database systems with fully replicated files: A hierarchical approach, *Performance Evaluation* **8**(3): 223–238.

Takagi, H. (1991). *Queueing Analysis; A Foundation of Performance Evaluation. Volume 1: Vacation and Priority Systems, Part 1*, Elsevier Science Publishers, Amsterdam.

Takine, T. and Sengupta, B. (1997). A single server queue with service interruptions, *Queueing Systems* **26**(3–4): 285–300.

Tang, Y. (1997). A single-server $M/G/1$ queueing system subject to breakdowns—some reliability and queueing problems, *Microelectonics Reliability* **37**(2): 315–321.

Tang, Y., Yun, X. and Huang, S. (2008). Discrete-time $Geo^X/G/1$ queue with unreliable server and multiple adaptive delayed vacations, *Journal of Computational and Applied Mathematics* **220**(1–2): 439–455.

Thiruvengadam, K. (1963). Queuing with breakdowns, *Operations Research* **11**(1): 62–71.

Tian, N. and Zhang, Z. (2006). *Vacation Queueing Models. Theory and Applications*, Springer, New York, NY.

Walraevens, J., Fiems, D. and Bruneel, H. (2006). The discrete-time preemptive repeat identical priority queue, *Queueing Systems* **53**(4): 231–243.

Walraevens, J., Steyaert, B. and Bruneel, H. (2004). Performance analysis of a $GI - Geo - 1$ buffer with a preemptive resume priority scheduling discipline, *European Journal of Operational Research* **157**(1): 130–151.

Wang, J.T. (2004). An $M/G/1$ queue with second optional service and server breakdowns, *Computers & Mathematics with Applications* **47**(10–11): 1713–1723.

Wang, L., Wang, C. and Adachi, F. (2011). Load-balancing spectrum decision for cognitive radio networks, *IEEE Journal on Selected Areas in Communications* **29**(4): 757–769.

White, H. and Christie, L. (1958). Queuing with preemptive priorities or with breakdown, *Operations Research* **6**(1): 79–95.

Woźniak, M., Kempa, W.M., Gabryel, M. and Nowicki, R.K. (2014). A finite-buffer queue with a single vacation policy: An analytical study with evolutionary positioning, *International Journal of Applied Mathematics and Computer Science* **24**(4): 887–900, DOI: 10.2478/amcs-2014-0065.

Wu, K., McGinnis, L. and Zwart, B. (2011). Queueing models for a single machine subject to multiple types of interruptions, *IIE Transactions* **43**(10): 753–759.

Yoon, C. and Un, C. (1994). Unslotted 1- and p_i-persistent CSMA-CD protocols for fiber optic bus networks, *IEEE Transactions on Communications* **42**(2–4): 158–465.

Zhang, F. and Zhu, Z. (2013). A discrete-time $Geo/G/1$ retrial queue with vacations and two types of breakdowns, *Journal of Applied Mathematics* **2013**, Article ID: 834731.

Dieter Fiems received the MSc degree in industrial engineering from KAHO-St-Lieven in 1997, the post-graduate degree in computer science from Ghent University in 1998, and the PhD degree in engineering from Ghent University in 2004. He is an associate professor at Ghent University, Department of Telecommunications and Information Processing. His current research interests include, amongst others, various applications of stochastic processes for the performance analysis of communication networks. Particularly, he is interested in applications of queueing theory and branching processes in wireless, optical and social networks.

Stijn De Vuyst was born in 1973. He obtained the MSc degree in electrical engineering in 1997 and the PhD degree in engineering sciences from Ghent University, Belgium, in 2005. He then became a post-doctoral researcher in the Department of Telecommunications and Information Processing, and in 2012 an assistant professor in the Department of Industrial Systems Engineering and Product Design at that university. His expertise is in operations research, in particular, stochastic modelling, simulation, queueing theory and scheduling with application to the design, planning and performance evaluation of production systems as well as telecommunication systems.

Received: 23 January 2018
Revised: 29 May 2018
Accepted: 20 June 2018