Queues with random back-offs

N. Bouman · S.C. Borst · O.J. Boxma · J.S.H. van Leeuwaarden

Received: 21 August 2012 / Revised: 17 May 2013 / Published online: 10 October 2013
© Springer Science+Business Media New York 2013

Abstract We consider a broad class of queueing models with random state-dependent vacation periods, which arise in the analysis of queue-based back-off algorithms in wireless random-access networks. In contrast to conventional models, the vacation periods may be initiated after each service completion, and can be randomly terminated with certain probabilities that depend on the queue length. We first present exact queue-length and delay results for some specific cases and we derive stochastic bounds for a much richer set of scenarios. Using these, together with stochastic relations between systems with different vacation disciplines, we examine the scaled queue length and delay in a heavy-traffic regime, and demonstrate a sharp trichotomy, depending on how the activation rate and vacation probability behave as function of the queue length. In particular, the effect of the vacation periods may either (i) completely vanish in heavy-traffic conditions, (ii) contribute an additional term to the queue lengths and delays of similar magnitude, or even (iii) give rise to an order-of-magnitude increase. The heavy-traffic trichotomy provides valuable insight into the impact of the back-off algorithms on the delay performance in wireless random-access networks.

Keywords Vacation queue · State-dependent vacations · Heavy-traffic analysis · CSMA protocol · Delay performance

Mathematics Subject Classification 60K25 · 68M20 · 90B22

N. Bouman (✉) · S.C. Borst · O.J. Boxma · J.S.H. van Leeuwaarden
Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: n.bouman@tue.nl

S.C. Borst
Alcatel-Lucent Bell Labs, P.O. Box 636, Murray Hill, NJ 07974-0636, USA
1 Introduction

We consider a broad class of queueing models with random state-dependent vacation periods. In contrast to conventional vacation models (see for instance [25] for a comprehensive overview), the server may take a vacation after each service completion and return from a vacation with certain probabilities that depend on the queue length. Specifically, when there are $i$ customers left behind in the system after a service completion, the server either takes a vacation with probability $\psi(i)$, with $\psi(0) \equiv 1$, or starts the service of the next customer otherwise. Likewise, the server returns from a vacation and starts the service of the next customer at the first event of a non-homogeneous Poisson process of rate $f(i)$, with $f(0) \equiv 0$, when there are $i$ customers in the system.

In view of the vacation discipline, we analyze the queue length process at departure epochs, unlike most papers on vacation models which consider the queue length process embedded at instants when vacations begin or end. A notable exception is [12], which studies an M/G/1 queue with a similar state-dependent vacation discipline, and establishes a stochastic decomposition property under certain assumptions. We show that this decomposition property in fact holds in far greater generality and corresponds to the Fuhrmann–Cooper decomposition. In addition, we obtain the exact stationary distributions of the queue length and delay for M/G/1 queues in three scenarios: (i) the probability $\psi(\cdot)$ decays geometrically as a function of the queue length, and the vacation is independent of the queue length; (ii) the probability $\psi(\cdot)$ is inversely proportional to the queue length, and the vacation is independent of the queue length; (iii) $\psi(\cdot) \equiv 1$ and the activation rate $f(\cdot)$ is proportional to the queue length.

We further derive lower and upper bounds for the mean queue length and mean delay in two cases: the activation rate $f(\cdot)$ is fixed and the vacation probability $\psi(\cdot)$ is a convex decreasing function; the activation rate $f(\cdot)$ is a concave or convex increasing function and $\psi(\cdot) \equiv 1$. Various stochastic bounds and comparison results are established as well.

We leverage the various bounds and stochastic comparison results to obtain the limiting distribution of the scaled queue length and delay in heavy-traffic conditions. The heavy-traffic asymptotics exhibit a sharp trichotomy. The first heavy-traffic regime emerges in scenarios (ii) and (iii) described above. In this regime the scaled queue length and delay converge to random variables with a gamma distribution. The commonality between these two scenarios lies in the fact that the ratio $f(i)/\psi(i)$ of the activation rate and the vacation probability is linear in the queue length. Loosely speaking, this means that the amount of vacation time is inversely proportional to the queue length. This proportionality property also holds for polling systems and in particular vacation queues with so-called branching type service disciplines, where the number of customers served in between two vacations (switch-over periods) is proportional to the queue length at the start of the service period. Interestingly, the scaled queue length and delay for these types of service disciplines have been proven to converge to random variables with a gamma distribution in heavy-traffic conditions as well. The significance of the ratio $f(i)/\psi(i)$ may also be recognized when the activation rate and vacation probability are in fact constant, i.e., independent of the queue length. In that case, the server is active a fixed fraction of the time which only depends on the activation rate and vacation probability through their ratio.
In the second heavy-traffic regime, which emerges in scenario (i) described above, the scaled queue length and delay both converge to an exponentially distributed random variable with the same mean as in the corresponding ordinary M/G/1 queue without any vacations. In other words, the impact of the vacations completely vanishes in heavy-traffic conditions. Note that in this scenario the vacation probability falls off faster than the inverse of the queue length.

The third heavy-traffic regime manifests itself when the vacation probability decays slower than the inverse of the queue length, e.g., like the inverse of the queue length raised to a power less than one. In that case, the queue length and delay, scaled by their respective means, converge to one in distribution, while the mean values increase an order-of-magnitude faster with the traffic intensity than in the first two regimes.

While the above results are of independent interest from a queueing perspective, they are also particularly relevant for the analysis of distributed medium access control algorithms in wireless networks, which in fact was the main motivation for the present work. Emerging wireless networks typically lack any centralized control entity for regulating transmissions, and instead vitally rely on the individual nodes to operate autonomously and fairly share the medium in a distributed fashion. A particularly popular mechanism for distributed medium access control is provided by the carrier-sense multiple-access (CSMA) protocol. In the CSMA protocol each node attempts to access the medium after a certain back-off time, but nodes that sense activity of interfering nodes freeze their back-off timer until the medium is sensed idle. From a queueing perspective, the back-off times may be interpreted as vacation periods during which no transmissions take place, even when packets may be queued up.

In the CSMA protocol, several powerful algorithms have been devised for adapting the back-off probabilities based on the queue lengths in non-saturated scenarios [13,21,24]. Roughly speaking, these algorithms provide maximum-stability guarantees under the condition that the back-off probabilities of the various nodes are reciprocal to the logarithms of the queue lengths. Unfortunately, however, such back-off probabilities can induce excessive queue lengths and delays, which has triggered a strong interest in developing approaches for improving the delay performance [2,3,5,10,14,18,19,23]. In particular, it has been shown that more aggressive schemes, where the back-off probabilities decay faster to zero as function of the queue lengths, can reduce the delays. The heavy-traffic results described above offer a useful indication of the impact of the choice of the back-off probabilities on the delay performance. It is worth observing that the vacation model does not account for the effects of the network topology, and highly aggressive schemes which are optimal in a single-node scenario, may in fact fail to achieve maximum stability in certain types of topologies [11]. However, the single-node results provide fundamental insight into how back-off probabilities may inherently inflate queue lengths and delays.

The remainder of the paper is organized as follows. In Sect. 2 we present a detailed model description. We provide an exact analysis of the model in Sect. 3, which yields formulas for the stationary queue length distribution in some specific cases. In Sect. 4 we derive lower and upper bounds for the mean queue length and we establish stochastic relations between systems with different functions $\psi(\cdot)$ and $f(\cdot)$. We study heavy-traffic behavior in Sect. 5 and identify three qualitatively different regimes. In Sect. 6 we summarize our findings and discuss the implications for wireless networks.
2 Model description

We consider an M/G/1 queue with vacations. That is, we consider a queueing system with one server that can be active or inactive. Customers arrive according to a Poisson process with rate \( \lambda \), independent of the state of the system. Let \( \sigma(t) \) indicate whether the server is active at time \( t \) (\( \sigma(t) = 1 \)) or not (\( \sigma(t) = 0 \)) and denote by \( L(t) \) the number of customers in the system at time \( t \).

When active, customers are served and the service times, generically denoted by \( B \), are generally distributed with distribution function \( F_B(\cdot) \) and Laplace–Stieltjes transform \( \tilde{B}(\cdot) \). We assume that \( \mathbb{E}\{B^2\} < \infty \) and that the service times are independent of the arrival and vacation times. Right after a service completion that leaves \( i \) customers behind, the server becomes inactive with probability \( \psi(i) \), where \( \psi : [0, \infty) \mapsto [0, 1] \).

Further we assume \( \psi(0) = 1 \), i.e., the server always becomes inactive if no customers are left in the system.

When inactive, no customer is served and we say that the server is on vacation. The server becomes active after some time that may depend on the number of waiting customers at the beginning of the vacation period and the number of customers that arrive during the vacation period, but it may not depend on future arrivals. Denote by \( \phi(i, m) \) the probability that exactly \( m \) customers arrive during a vacation period that begins with \( i \) customers in the system, where \( \phi : [0, \infty) \times [0, \infty) \mapsto [0, 1] \). Further we assume \( \phi(0, 0) = 0 \), i.e., the server does not activate if no customers are present in the system. Let the random variable \( X_i \) denote the number of arrivals during a vacation period that begins with \( i \) customers in the system. We will assume that \( \phi(\cdot) \) is such that \( \mathbb{E}\{X_i^2\} < \infty \).

In the literature on CSMA-like queue-based random access algorithms it is usually assumed that inactive transmitters (or servers) activate according to some Poisson process whenever they are not blocked by their neighbors. For an isolated transmitter (or server) this corresponds to assuming that the transmitter (or server) becomes active at the first jump of a non-homogeneous Poisson process with rate \( f(i) \) when \( L(i) = i \), i.e., taking

\[
\phi(i, m) = \frac{f(i + m)}{\lambda + f(i + m)} \prod_{j=0}^{m-1} \frac{\lambda}{\lambda + f(i + j)}
\]

in our model. Further the transmission (or service) times are usually assumed to be exponentially distributed, so that \( \psi(i)/\mathbb{E}\{B\} \) is the exponential de-activation rate of an active transmitter (or server) when \( L(t) = i \). This queue-based random access algorithm is shown to be maximally stable in any network topology for certain \( f(\cdot) \) and \( \psi(\cdot) \) with \( f(i)/\psi(i) \sim \log(i) \) [13,21,24].

Let \( \rho = \lambda \mathbb{E}\{B\} \) denote the traffic intensity of the system. Throughout this paper, we denote the generating function of a non-negative and discrete random variable \( W \) by \( G_W(r) = \mathbb{E}\{r^W\} \), with \( r \in [0, 1] \). Note that

\[
G_{X_i}(r) = \sum_{m=0}^{\infty} \phi(i, m) r^m.
\]
Let \( W_1 \overset{d}{=} W_2 \) denote that two random variables \( W_1 \) and \( W_2 \) are equal in distribution, so that \( \mathbb{P}\{W_1 \leq w\} = \mathbb{P}\{W_2 \leq w\} \) for all \( w \). Further, let \( W_1 \geq_{st} W_2 \) denote that \( W_1 \) is stochastically larger than \( W_2 \), so that \( \mathbb{P}\{W_1 \leq w\} \leq \mathbb{P}\{W_2 \leq w\} \) for all \( w \). Finally, let \( W_1 >_{st} W_2 \) denote that \( W_1 \) is stochastically strictly larger than \( W_2 \), so that \( W_1 \geq_{st} W_2 \) and, additionally, \( \mathbb{P}\{W_1 \leq w\} < \mathbb{P}\{W_2 \leq w\} \) for some \( w \).

### 3 Exact analysis

Denote by \( Z_n \) the number of customers just after the \( n \)th service completion, i.e., the number of customers left behind by the \( n \)th departing customer. Further denote by \( A_n \) the number of arrivals during the \( n \)th service and note that \( A_n \) does not depend on \( X_i \).

Then \( (Z_n)_{n \in \mathbb{N}_0} \) constitutes a Markov chain with transition probabilities

\[
\mathbb{P}\{Z_{n+1} = j | Z_n = i\} = (1 - \psi(i))\mathbb{P}\{A_n = j - i + 1\} + \psi(i)\mathbb{P}\{X_i + A_n = j - i + 1\},
\]

for \( j \geq i - 1 \) and

\[
\mathbb{P}\{Z_{n+1} = j | Z_n = i\} = 0,
\]

for \( j < i - 1 \). Because \( X_i \) and \( A_n \) are independent,

\[
\mathbb{E}\{r^{Z_{n+1}} | Z_n = i\} = (1 - \psi(i))r^{i-1}G_{A_n}(r) + \psi(i)r^{i-1}G_{X_i}(r)G_{A_n}(r) = r^{i-1}G_{A_n}(r)(1 + \psi(i)(G_{X_i}(r) - 1)),
\]

where \( G_{A_n}(r) = \hat{B}(\lambda(1 - r)) \) for all \( n \). Using this relation we can find a sufficient condition for stability of the system.

**Lemma 3.1** The Markov chain \( (Z_n)_{n \in \mathbb{N}_0} \) is positive recurrent if

\[
\limsup_{i \to \infty} \psi(i)\mathbb{E}\{X_i\} < 1 - \rho. \tag{3}
\]

**Proof** This result is proved in [12], using the results in [8]. For a short proof note that from (2) we find \( \mathbb{E}\{Z_{n+1} | Z_n = i\} = i - 1 + \rho + \psi(i)\mathbb{E}\{X_i\} \). The result now follows immediately from Pakes’ Lemma [20] as \( \mathbb{E}\{X_i\} < \infty \) for all \( i \geq 0 \) by assumption.

\[\square\]

In words, Lemma 3.1 states that for stability it is sufficient that the system is busy serving customers more than a fraction \( \rho \) of the time if the number of customers in the system is large.

We henceforth assume the system is stable, i.e., \( \psi(\cdot) \) and \( \phi(\cdot) \) are such that the condition in (3) is satisfied. Let the random variable \( Z \) have the stationary distribution of the embedded Markov chain \( (Z_n)_{n \in \mathbb{N}_0} \), i.e.,

\[
\mathbb{P}\{Z = j\} = \lim_{n \to \infty} \mathbb{P}\{Z_n = j | Z_0 = k\}, \quad k \geq 0.
\]
By the PASTA property and a level crossings argument we know that \( \mathbb{P}\{Z = j\}, j \geq 0 \) is also the stationary distribution of the number of customers in the system \( L \), with
\[
\mathbb{P}\{L = j\} = \lim_{t \to \infty} \mathbb{P}\{L(t) = j|L(0) = k\},
\]
for any \( k \geq 0 \). Hence, using (2), we obtain the relation
\[
G_L(r) = \frac{\tilde{B}(\lambda(1 - r))}{\tilde{B}(\lambda(1 - r)) - r} \left( G_L(r) + \sum_{i=0}^{\infty} \psi(i)r^i \mathbb{P}\{L = i\}(G_{X_i}(r) - 1) \right),
\]
which corresponds to [12, Eq. (2)]. Equivalently,
\[
G_L(r) = \frac{\tilde{B}(\lambda(1 - r)) \sum_{i=0}^{\infty} \psi(i)r^i \mathbb{P}\{L = i\}(1 - G_{X_i}(r))}{\tilde{B}(\lambda(1 - r)) - r}.
\]
(4)

**Example 3.1** One activation scheme would be to never de-activate when the system is nonempty right after a service completion, i.e., \( \psi(i) = 0 \) for \( i \geq 1 \). Similarly we could say that the server always activates immediately if there are waiting customers at the beginning of the vacation period, i.e., \( \phi(i, 0) = 1 \) for \( i \geq 1 \), and \( \phi(i, m) = 0 \) otherwise, so that \( G_{X_i}(r) = 1 \) for \( i \geq 1 \). For this activation scheme (4) simplifies to
\[
G_L(r) = \frac{\tilde{B}(\lambda(1 - r)) \mathbb{P}\{L = 0\}(1 - G_{X_0}(r))}{\tilde{B}(\lambda(1 - r)) - r}.
\]
Using that \( G_L(1) = 1 \) and applying l’Hôpital’s rule yields
\[
\mathbb{P}\{L = 0\} = \frac{1 - \rho}{\mathbb{E}\{X_0\}},
\]
and hence
\[
G_L(r) = \frac{(1 - \rho)\tilde{B}(\lambda(1 - r))(1 - G_{X_0}(r))}{\mathbb{E}\{X_0\}(\tilde{B}(\lambda(1 - r)) - r)}.
\]
(5)

Note that if the server waits for exactly one customer to arrive if there are no waiting customers at the beginning of the vacation period, i.e., \( X_0 \equiv 1 \), then (5) becomes the classical Pollaczek–Khinchin formula for the standard M/G/1 queue without vacations,
\[
G_L(r) = G_{LM/G/1}(r) = \frac{(1 - \rho)\tilde{B}(\lambda(1 - r))(1 - r)}{\tilde{B}(\lambda(1 - r)) - r}.
\]
(6)

The Fuhrmann–Cooper decomposition [9] relates \( G_L(r) \) to the Pollaczek–Khinchin formula through

\[ \square \]
GL(r) = G_{LM/G/1}(r)G_{L_1}(r),  \tag{7}

where \( L_1 \) denotes the number of customers in the system at an arbitrary epoch during a non-serving (vacation) period. This decomposition property can be derived from (4). For this denote by \( L_{\text{begin}} \) and \( L_{\text{end}} \) the number of customers in the system at, respectively, the beginning and the end of a vacation period, and let \( \gamma \) be the probability that the server becomes inactive after a departure,

\[
\gamma = \sum_{i=0}^{\infty} \psi(i) \mathbb{P}\{L = i\}.
\]

Because the system is stable the expected number of arrivals between two service completions is equal to the expected number of service completions, which equals one. Therefore,

\[
\rho + \gamma (\mathbb{E}\{L_{\text{end}}\} - \mathbb{E}\{L_{\text{begin}}\}) = 1,
\]

and the expected number of arrivals during a vacation period is therefore given by

\[
\mathbb{E}\{L_{\text{end}}\} - \mathbb{E}\{L_{\text{begin}}\} = \frac{1 - \rho}{\gamma}.
\]

Further note that

\[
\mathbb{P}\{L_{\text{begin}} = i\} = \frac{1}{\gamma} \mathbb{P}\{L = i\} \psi(i).
\]

From (4) we now find

\[
G_L(r) = \frac{(1 - \rho) \tilde{B}(\lambda(1 - r))(1 - r) \sum_{i=0}^{\infty} \psi(i) r^i \mathbb{P}\{L = i\} (1 - G_{X_1}(r))}{\tilde{B}(\lambda(1 - r)) - r (1 - \rho)(1 - r)}
\]

\[
= G_{LM/G/1}(r) \sum_{i=0}^{\infty} \frac{1}{\gamma} \psi(i) r^i \mathbb{P}\{L = i\} (1 - G_{X_1}(r))
\]

\[
= G_{LM/G/1}(r) \sum_{i=0}^{\infty} \frac{r^i \mathbb{P}\{L_{\text{begin}} = i\} (1 - G_{X_1}(r))}{(1 - r)(\mathbb{E}\{L_{\text{end}}\} - \mathbb{E}\{L_{\text{begin}}\})}
\]

\[
= G_{LM/G/1}(r) \frac{G_{L_{\text{begin}}}(r) - G_{L_{\text{end}}}(r)}{(1 - r)(\mathbb{E}\{L_{\text{end}}\} - \mathbb{E}\{L_{\text{begin}}\})},
\]

yielding (7); see [1]. Thus to find \( G_L(r) \) we can either solve Eq. (4) or find \( G_{L_1}(r) \) and then use the Fuhrmann–Cooper decomposition.

In the remainder of this section we will analyze the system for several choices of \( \phi(\cdot) \) and \( \psi(\cdot) \).
3.1 Equal vacation distributions

In this subsection we assume that \( X_i \overset{d}{=} X \) for \( i \geq 1 \), with \( X \) some generic random variable. We further assume that \( X \overset{d}{=} X \) for \( i \geq 1 \), with \( X \) some generic random variable. We further assume that \( X >_{st} 0 \), so that with nonzero probability at least one customer arrives during any vacation. The case \( X \overset{d}{=} 0 \) is already solved in Example 3.1.

Next, if a vacation starts with no customers in the system, we assume that first \( X \) customers arrive. After this, if the system is still empty, the vacation is extended in an arbitrary way until at least one customer has arrived. We thus have \( X_0 \overset{d}{=} X \) if \( X \geq_{st} 1 \), i.e., if \( \mathbb{P}\{X = 0\} = 0 \), and \( X_0 >_{st} X \) otherwise, i.e., if \( \mathbb{P}\{X = 0\} > 0 \).

To summarize, in this subsection we study the following scenario.

**Scenario 1** \( X_i \overset{d}{=} X >_{st} 0 \) for all \( i \geq 1 \) and, either \( X_0 \overset{d}{=} X \) and \( \mathbb{P}\{X = 0\} = 0 \), or \( X_0 >_{st} X \) and \( \mathbb{P}\{X = 0\} > 0 \).

Note that in this scenario we have \( G_{X_i}(r) = G_X(r) \) for all \( i \geq 0 \) and all \( r \in [0, 1] \) if \( G_X(0) = 0 \), and \( G_{X_0}(r) < G_{X_i}(r) = G_X(r) \) for all \( i \geq 1 \) and all \( r \in [0, 1] \) if \( G_X(0) > 0 \). So from (4) and \( \psi(0) = 1 \) it follows that

\[
G_L(r) = \frac{\tilde{B}(\lambda(1 - r))\left(\mathbb{P}\{L = 0\}(G_X(r) - G_{X_0}(r)) + (1 - G_X(r)) \sum_{i=0}^{\infty} \psi(i)r^i\mathbb{P}\{L = i\}\right)}{\tilde{B}(\lambda(1 - r)) - r}.
\]

Equation (8) seems hard to solve in general, but we are able to find solutions for several specific choices for \( \psi(\cdot) \). Before analyzing (8) in more detail we now first give a prototypical example of a system that belongs to Scenario 1. This example describes a back-off mechanism used in wireless networks.

**Example 3.2** Consider a server that always waits for a certain time \( V \), independent of the arrivals during this time. After this time the server activates if there are customers present in the system, and otherwise the server again waits for a time \( V \) (independent of the previous time) and repeats this procedure until there are customers present in the system. Assume \( V \) is generally distributed with distribution function \( F_V(\cdot) \) and Laplace–Stieltjes transform \( \tilde{V}(\cdot) \). Denoting by \( \alpha_m \) the probability that exactly \( m \) customers arrive during a time \( V \),

\[
\alpha_m = \int_{0}^{\infty} \frac{(\lambda t)^m}{m!} e^{-\lambda t} dF_V(t),
\]

we have \( \phi(i, m) = \alpha_m \), for \( i \geq 1 \), and \( \phi(0, m) = \alpha_m / (1 - \alpha_0) \), for \( m \geq 1 \). Further, we get

\[
G_{X_0}(r) = \sum_{m=1}^{\infty} \phi(0, m)r^m = \frac{1}{1 - \alpha_0} \int_{t=0}^{\infty} \sum_{m=1}^{\infty} \frac{(\lambda tr)^m}{m!} e^{-\lambda t} dF_V(t) = \frac{\tilde{V}(\lambda(1 - r)) - \tilde{V}(\lambda)}{1 - \tilde{V}(\lambda)}.
\]
where the interchange of summation and integration is justified by Beppo Levi’s theorem, see for example [7]. Similarly, we get $G_{X_i}(r) = \tilde{V}(\lambda(1-r))$ for $i \geq 1$.

Note that if $V$ is exponentially distributed with mean $1/\nu$ we find $X_i^{\text{d}} = X$, $i \geq 1$, where $X$ is a geometric random variable, with

$$G_X(r) = \frac{\nu}{\lambda(1-r) + \nu}.$$  \hspace{1cm} (9)

Further, $G_{X_0}(r) = rG_{X_1}(r)$ as $X_0^{\text{d}} = X_1 + 1$ in this case.

We will now use (8) to find $G_L(\cdot)$ if $\psi(i) = ai$ or $\psi(i) = 1/(i+1)$, two functions that we will need in the heavy-traffic analysis of the system, see Sect. 5. For this purpose first introduce

$$K(r) = \frac{\tilde{B}(\lambda)(1-r)(G_X(r) - G_{X_0}(r))}{\tilde{B}((1-r)) - r},$$

with $K(r) \equiv 0$ if $X_0^{\text{d}} = X$. Define

$$Y(r) = \frac{\tilde{B}(\lambda)(1-r)X(r)}{\tilde{B}((1-r)) - r},$$

and note that an alternative expression for $Y(\cdot)$ is given by

$$Y(r) = G_{LM/G/1}(r)G_{X^{\text{res}}}(r)\frac{\mathbb{E}[X]}{1-\rho},$$

with $G_{LM/G/1}(r)$ the generating function of the number of customers in a standard M/G/1 queue without vacations as in (6), and $G_{X^{\text{res}}}(r)$ the generating function of the number of arrivals in a residual vacation period,

$$G_{X^{\text{res}}}(r) = \mathbb{E}[r^{X^{\text{res}}}] = \frac{1 - G_X(r)}{(1-r)\mathbb{E}[X]}.$$  \hspace{1cm} (12)

Similarly, we can write

$$K(r) = G_{LM/G/1}(r)\frac{G_X(r) - G_{X_0}(r)}{(1-\rho)(1-r)}$$

$$= G_{LM/G/1}(r)\left(\mathbb{E}[X_0]G_{X^{\text{res}}}(r) - \mathbb{E}[X]G_{X^{\text{res}}}(r)\right)\frac{1}{1-\rho}.$$ \hspace{1cm} (13)

Further, by l’Hôpital’s rule,

$$Y(1) = \lim_{r \uparrow 1} Y(r) = \frac{\mathbb{E}[X]}{1-\rho}.$$
and

\[ K(1) = \lim_{r \uparrow 1} K(r) = \frac{\mathbb{E}[X_0] - \mathbb{E}[X]}{1 - \rho}. \]

Note that \( Y(1) > 0 \) as \( \mathbb{E}[X] > 0 \) and \( \rho < 1 \) for stability. Also note that \( K(1) > 0 \) if \( X_0 >_{st} X \).

Finally note that the generating function \( G_W(r) \) of any non-negative discrete random variable \( W \) is a non-negative continuously differentiable function on \([0, 1]\), as follows from the definition of a generating function. Hence \( Y(r) \) and \( K(r) \) are non-negative continuously differentiable functions on \([0, 1]\).

**Theorem 3.2** Consider Scenario 1 and \( \psi(i) = a^i \) with \( 0 \leq a < 1 \), \( i \geq 0 \).

(i) If \( X_0 \overset{d}{=} X \), then

\[ G_L(r) = \frac{\prod_{i=0}^{\infty} Y(a^i r)}{\prod_{i=0}^{\infty} Y(a^i)} \]  \hspace{1cm} (14)

(ii) If \( X_0 >_{st} X \), then

\[ G_L(r) = \frac{\sum_{j=0}^{\infty} K(a^j r) \prod_{i=0}^{j-1} Y(a^i r)}{\sum_{j=0}^{\infty} K(a^j) \prod_{i=0}^{j-1} Y(a^i)}, \]  \hspace{1cm} (15)

with \( \prod_{i=0}^{j-1} g(i) = 1 \) for any function \( g(\cdot) \).

**Proof** From Lemma 3.1 we obtain that the system is stable if \( \rho < 1 \), as \( 0 \leq a < 1 \). We will now first prove the result for case (i), for which (8) simplifies to

\[ G_L(r) = \frac{\tilde{B}(\lambda(1 - r))(1 - G_X(r))G_L(ar)}{\tilde{B}(\lambda(1 - r)) - r} = Y(r)G_L(ar). \]  \hspace{1cm} (16)

Upon iteration this gives, using \( G_L(0) = \mathbb{P}[L = 0] \),

\[ G_L(r) = \mathbb{P}[L = 0] \prod_{i=0}^{\infty} Y(a^i r). \]  \hspace{1cm} (17)

Finally, using \( G_L(1) = 1 \), we obtain

\[ \mathbb{P}[L = 0] = \frac{1}{\prod_{i=0}^{\infty} Y(a^i)}. \]  \hspace{1cm} (18)

Combining (17) and (18) yields assertion (14). In Lemma 7.1 we prove that \( \prod_{i=0}^{\infty} Y(a^i r) \) converges for all \( r \in [0, 1] \), so in particular \( \mathbb{P}[L = 0] > 0 \).
For case (ii) Eq. (8) simplifies to

$$GL(r) = \frac{\tilde{B}(\lambda(1-r))(P[L = 0](G_X(r) - G_{X_0}(r)) + (1 - G_X(r))G_L(ar))}{\tilde{B}(\lambda(1-r)) - r} = K(r)P[L = 0] + Y(r)G_L(ar).$$

Iterating this and using $G_L(0) = P[L = 0]$ we get

$$GL(r) = P[L = 0]\left(\sum_{j=0}^{\infty} K(a^j r) \prod_{i=0}^{j-1} Y(a^i r) + \prod_{i=0}^{\infty} Y(a^i r)\right).$$

Now note that $Y(0) = 1 - G_X(0)$, so $Y(0) < 1$ by assumption. Thus, as $Y(\cdot)$ is continuous and $0 \leq a < 1$,

$$GL(r) = P[L = 0]\sum_{j=0}^{\infty} K(a^j r) \prod_{i=0}^{j-1} Y(a^i r).$$

From (19) and $G_L(1) = 1$ we get (15). In Lemma 7.1 we prove that $\sum_{j=0}^{\infty} K(a^j r) \prod_{i=0}^{j-1} Y(a^i r)$ converges for all $r \in [0, 1]$, so in particular

$$P[L = 0] = \frac{1}{\sum_{j=0}^{\infty} K(a^j) \prod_{i=0}^{j-1} Y(a^i)} > 0,$$

which completes the proof. \(\Box\)

**Theorem 3.3** Consider Scenario 1 and $\psi(i) = 1/(i + 1)$, $i \geq 0$, with $\alpha(r) = \int_r^1 Y(x) \frac{1}{x} dx$.

(i) If $X_0 \overset{d}{=} X$, then

$$GL(r) = \frac{(1 - \rho)Y(r)}{rE[X]} e^{-\alpha(r)}.$$

(ii) If $X_0 \overset{st}{=} X$, then

$$GL(r) = P[L = 0] \left(K(r) + \frac{Y(r)}{r} e^{-\alpha(r)} \int_0^r K(y) e^{\alpha(y)} dy\right),$$

with

$$P[L = 0] = \frac{1}{K(1) + Y(1) \int_0^1 K(x) e^{\alpha(x)} dx}.$$
Proof From Lemma 3.1 it follows that the system is stable if $\rho < 1$ as $\lim_{i \to \infty} \psi(i) = 0$. For case (i), Eq. (8) gives

$$G_L(r) = \frac{\tilde{B}(\lambda(1 - r))(1 - G_X(r)) \sum_{i=0}^{\infty} \frac{1}{i+1} r^i \mathbb{P}[L = i]}{\tilde{B}(\lambda(1 - r)) - r}.$$ 

Now define

$$H_L(r) = \sum_{i=0}^{\infty} \frac{1}{i+1} r^{i+1} \mathbb{P}[L = i],$$

and note that $H_L'(r) = G_L(r)$. Thus, $H_L(r)$ can be found by solving the differential equation

$$H_L'(r) = \frac{Y(r)H_L(r)}{r}. \quad (23)$$

We thus get

$$H_L(r) = C \cdot \exp \left( - \int \frac{1}{r} \frac{Y(x)}{x} \, dx \right)$$

and

$$G_L(r) = C \cdot \frac{Y(r)}{r} \exp \left( - \int \frac{1}{r} \frac{Y(x)}{x} \, dx \right),$$

where $C$ is some constant. Using $G_L(1) = 1$ this gives (20).

For case (ii) we can find $H_L(r)$ by solving the differential equation

$$H_L'(r) = \mathbb{P}[L = 0] K(r) + \frac{Y(r)H_L(r)}{r}. \quad (24)$$

This gives

$$H_L(r) = e^{-\alpha(r)} \left( \mathbb{P}[L = 0] \int_0^r K(y) e^{\alpha(y)} \, dy + C \right)$$

and

$$G_L(r) = \mathbb{P}[L = 0] K(r) + \frac{Y(r)}{r} e^{-\alpha(r)} \left( \mathbb{P}[L = 0] \int_0^r K(y) e^{\alpha(y)} \, dy + C \right), \quad (25)$$
for some constant $C$. As $G_L(r)$ is a generating function, boundary conditions are given by $G_L(0) = \mathbb{P}\{L = 0\}$ and $G_L(1) = 1$.

To solve this boundary problem, note that, using integration by parts,

$$\alpha(r) = - \log(r) Y(r) - \int_r^1 Y'(x) \log(x) \, dx.$$  

Further $Y(0) = 1 - G_X(0)$, so $Y(0) < 1$ in this case, and for all $x \in [0, 1]$, $Y'(x) \leq Y'(1) < \infty$ as $\mathbb{E}\{B^2\} < \infty$ and $\mathbb{E}\{X^2\} < \infty$. Therefore,

$$\frac{Y(r)}{r} e^{-\alpha(r)} = Y(r) r^{Y(r)-1} \exp\left(\int_r^1 Y'(x) \log(x) \, dx\right)$$

and thus

$$\lim_{r \downarrow 0} \frac{Y(r)}{r} e^{-\alpha(r)} \geq \lim_{r \downarrow 0} Y(r) \exp\left((Y(r) - 1) \log(r)\right) \exp(-Y'(1)) = \infty.$$  

Hence, to get $G_L(0) = \mathbb{P}\{L = 0\}$, we need to set $C = 0$, so that (25) becomes

$$G_L(r) = \mathbb{P}\{L = 0\}\left(K(r) + \frac{Y(r)}{r} e^{-\alpha(r)} \int_0^r K(y) e^{\alpha(y)} \, dy\right).$$  

Finally, using $G_L(1) = 1$ we find (21).

Equation (8) can be used to find $G_L(\cdot)$ for other functions $\psi(\cdot)$ as well. For example, if $\psi(i) = g(i)$ for $i < I$ and $\psi(i) = a_i$ for $i \geq I$, for some function $g(\cdot), 0 \leq g(\cdot) \leq 1$, and some $I \in \mathbb{N}_0$, one can use the same approach as used in the proof of Theorem 3.2 to find $G_L(\cdot)$ in terms of $\mathbb{P}\{L = j\}, j = 0, \ldots, I - 1$. Although interesting, it is beyond the scope of this paper to analyze this in more detail.

3.2 State-dependent vacation lengths

In this subsection we consider a system that has state-dependent activation rules. More precisely, we assume that the server becomes active at the first jump of a non-homogeneous Poisson process with rate $f(i)$ when $L(t) = i$. This gives the following scenario, see also (1).

**Scenario 2**  

$$\phi(i, m) = \frac{f(i + m)}{\lambda + f(i + m)} \prod_{j=0}^{m-1} \frac{\lambda}{\lambda + f(i + j)}, \quad i, m \geq 0,$$

where $f : [0, \infty) \mapsto [0, \infty)$ and $f(0) = 0$.

Note that we need $f(0) = 0$ as $\phi(0, 0) = 0$ if and only if $f(0) = 0$.  

\(\copyright\) Springer
Theorem 3.4 Consider Scenario 2 with $\psi(i) = 1$ and $f(i) = vi$, $i \geq 0$. Then,

$$G_L(r) = \frac{(1 - \rho)\tilde{B}(\lambda(1 - r))(1 - r)}{\tilde{B}(\lambda(1 - r)) - r}\exp\left(\int_r^1 \frac{-\lambda(1 - x)}{\nu(\tilde{B}(\lambda(1 - x)) - x)}\,dx\right). \quad (26)$$

Proof First note that, for all $m > 0$, $\phi(i, m) \to 0$ as $i \to \infty$. So $E\{X_i\} \to 0$ and the system is stable if $\rho < 1$, see Lemma 3.1.

To prove this theorem we will first determine $G_{L_I}(r)$, and then $G_L(r)$ follows from the Fuhrmann–Cooper decomposition (7).

In order to determine the distribution of $L_I$ we cut out all the services and replace these by instantaneous jumps whose sizes are the number of arrivals $L_A$ during an arbitrary service time, with $G_{L_A}(r) = \tilde{B}(\lambda(1 - r))$. These jumps occur at rate $\nu i$ when there are $i$ customers in the system. Thus, as the function $f(\cdot)$ is linear, the distribution of $L_I$ corresponds to that of a continuous-time branching process with immigration: Particles arrive as a Poisson process with rate $\lambda$ and each particle is independently and instantaneously replaced at rate $\nu$ by a new group of $L_A$ particles. Branching processes of this type were studied by Sevast’yanov [22] and applying [22, Thm. 1] to our situation yields

$$G_{L_I}(r) = \exp\left(\int_r^1 \frac{-\lambda(1 - x)}{\nu(\tilde{B}(\lambda(1 - x)) - x)}\,dx\right).$$

By (7) we then obtain (26). \qed

3.2.1 Exponentially distributed service times

For generally distributed service times it is difficult to interpret Theorem 3.4, but for exponentially distributed service times we can using the following result.

Corollary 3.5 Consider Scenario 2 with $\psi(i) = 1$ and $f(i) = vi$, $i \geq 0$, and exponentially distributed service times with mean $1/\mu$. Then,

$$G_L(r) = \left(\frac{1 - \rho}{1 - \rho r}\right)^{1+\lambda/\nu} e^{(r-1)\lambda/\nu}. \quad (27)$$

Proof Evaluating the integral in Theorem 3.4 with $\tilde{B}(s) = \frac{\mu}{\mu + s}$ gives (27). \qed

Notice that (27) is the product of two generating functions, so that the distribution of $L$ is a convolution of a negative binomial distribution and a Poisson distribution.

For exponentially distributed service times, and any choice of $f(\cdot)$ and $\psi(\cdot)$, $(L(t), \sigma(t))_{t \geq 0}$ is a continuous-time Markov process with state space $\{0, 1, 2, \ldots\} \times \{0, 1\}$ and state $(i, j)$ representing $i$ customers in the system and $j = 0$ when the server is inactive and $j = 1$ when the server is active. Transitions from $(i, 0)$ to $(i + 1, 0)$ and from $(i, 1)$ to $(i + 1, 1)$ occur at rate $\lambda$, corresponding to an arrival of a customer, and
transitions from \((i, 0)\) to \((i, 1)\) occur at rate \(f(i)\), corresponding to a server activation. Further, transitions from \((i + 1, 1)\) to \((i, 0)\) occur at rate \(\mu \psi(i)\), corresponding to a service completion and server de-activation. Finally, transitions from \((i + 1, 1)\) to \((i, 1)\) occur at rate \(\mu(1 - \psi(i))\), corresponding to a service completion without server de-activation. With \(\pi(i, k)\) the stationary probability that the Markov process resides in state \((i, k)\), we have the balance equations

\[
\begin{align*}
\lambda \pi(0, 0) &= \mu \pi(1, 1), \\
(\lambda + \mu)\pi(1, 1) &= f(1)\pi(1, 0) + \mu(1 - \psi(1))\pi(2, 1), \\
(\lambda + f(i))\pi(i, 0) &= \lambda\pi(i - 1, 0) + \mu \psi(i)\pi(i + 1, 1), \quad i \geq 1, \\
(\lambda + \mu)\pi(i, 1) &= \lambda\pi(i - 1, 1) + f(i)\pi(i, 0) + \mu(1 - \psi(i))\pi(i + 1, 1), \quad i \geq 2.
\end{align*}
\]

This set of balance equations can be solved for several choices of \(f(\cdot)\) and \(\psi(\cdot)\). For example, \(f(i) = b^i\), with \(b > 1\), and \(\psi(i) = 1, i \geq 0\), yields a result similar to the result of Theorem 3.2. Also, the result of Corollary 3.5 can be derived in this way.

The next theorem gives a class of functions for which the distribution of the total number of customers in the system in steady state is negative binomial.

**Theorem 3.6** Consider Scenario 2 with \(\psi(i) = k/(k + i), i \geq 1\), and \(f(i) = \mu i/(i + k - 1), i \geq 0\), with \(k \geq 0\), and exponentially distributed service times with mean \(1/\mu\). Then,

\[
G_L(r) = \left( \frac{1 - \rho}{1 - \rho r} \right)^{k+1}.
\]  
(28)

**Proof** It can be checked that

\[
\pi(i, 0) = \binom{i + k - 1}{i} (1 - \rho)^{k+1} \rho^i,
\]

and

\[
\pi(i, 1) = \binom{i + k - 1}{i - 1} (1 - \rho)^{k+1} \rho^i,
\]

solve the set of balance equations and the normalization equation \(\sum_{i, j} \pi(i, j) = 1\). Thus \(\mathbb{P}(L = 0) = \pi(0, 0) = (1 - \rho)^{k+1}\) and for \(i \geq 1\),

\[
\mathbb{P}(L = i) = \pi(i, 0) + \pi(i, 1) = \binom{i + k}{i} (1 - \rho)^{k+1} \rho^i.
\]

Recognizing the probability mass function of the negative binomial distribution with stopping parameter \(k + 1\) we find (28). \(\square\)

Note that the functions in Theorem 3.6 describe an M/M/1 queue if \(k = 0\), as one always has an immediate transition from \((1, 0)\) to \((1, 1)\) and there are no transitions from \((i, 1)\) to \((k, 0)\) for any \(k \geq 0\) if \(i \geq 2\).
Further, \( k = 1 \) leads to a special case of Theorem 3.3, because \( \psi(i) = 1/(i + 1) \), \( i \geq 1 \), the service times are exponentially distributed with mean \( 1/\mu \) and the vacation discipline of Example 3.2 is used with the vacation time distribution identical to the service time distribution.

The results of Corollary 3.5 and Theorem 3.6 could also be derived using a probabilistic approach. For the situation of Corollary 3.5 the number of customers at an arbitrary epoch during a vacation period \( L_I \) can be related to the customers in a network of infinite-server queues with phase-type service requirement distributions.

For the situation of Theorem 3.6 the vacation model behaves as an M/M/1 queue with \( k \) permanent customers and a Random-Order-of-Service (ROS) discipline. The ROS discipline selects the next customer for service at random from those which were in the queue just before the service completion, and excludes a permanent customer whose service may just have been completed.

4 Bounds

In Sect. 3 we obtained exact results for several choices of \( \phi(\cdot) \) and \( \psi(\cdot) \). In this section we derive bounds for the mean number of customers in the system and we establish stochastic relations between systems with different activation schemes. These bounds and stochastic relations will be used in the heavy-traffic analysis in Sect. 5.

4.1 Equal vacation distributions

In this subsection we consider the class of vacation disciplines of Scenario 1 as described in Sect. 3.1. For this class we find the following lower bound.

**Theorem 4.1** Consider Scenario 1 and let \( \psi(\cdot) \) be a strictly decreasing convex function. Then,

\[
\mathbb{E}\{L\} \geq \max \left\{ \psi^{-1}\left(\frac{1 - \rho}{\mathbb{E}\{X\}}\right), \frac{\lambda^2\mathbb{E}\{B^2\}}{2(1 - \rho)} + \rho \right\}.
\]

**Proof** In steady state, the mean number of activations per unit of time equals the mean number of de-activations per unit of time, so that

\[
P[\sigma = 1] \frac{1}{\mathbb{E}\{B\}} \mathbb{E}\{\psi(Z)\} = \frac{\lambda}{\mathbb{E}\{X_0\}} P[\sigma = 0 \cap L_{\text{begin}} = 0] + \frac{\lambda}{\mathbb{E}\{X\}} P[\sigma = 0 \cap L_{\text{begin}} > 0],
\]

where \( \sigma \) denotes the random variable with the steady-state distribution of the state of the server, i.e.,

\[
P[\sigma = j] = \lim_{t \to \infty} P[\sigma(t) = j | \sigma(0) = k],
\]

and \( Z \) and \( L_{\text{begin}} \) denote, as before, respectively the steady-state number of customers in the system right after service completions and at the start of a vacation period.
Because \( \mathbb{E}(X_0) \geq \mathbb{E}(X) \), \( Z \overset{d}{=} L \) and \( \mathbb{P}(\sigma = 1) = \rho \), (29) gives
\[
\lambda \mathbb{E}(\psi(L)) \leq \frac{\lambda}{\mathbb{E}(X)} \mathbb{P}(\sigma = 0) = \frac{\lambda}{\mathbb{E}(X)} (1 - \rho).
\]
Further, it follows from Jensen’s inequality that, as \( \psi(\cdot) \) is convex,
\[
\mathbb{E}(\psi(L)) \geq \psi(\mathbb{E}(L)).
\]
Since \( \psi(\cdot) \) is decreasing we then get
\[
\mathbb{E}(L) \geq \psi^{-1}\left(\frac{1 - \rho}{\mathbb{E}(X)}\right).
\] (30)
Finally, the Fuhrmann–Cooper decomposition (7) implies
\[
\mathbb{E}(L) = \mathbb{E}(L_{M/G/1}) + \mathbb{E}(L_I) = \frac{\lambda^2 \mathbb{E}(B^2)}{2(1 - \rho)} + \rho + \mathbb{E}(L_I),
\]
where \( \mathbb{E}(L_{M/G/1}) \) follows from the Pollaczek–Khinchin formula (6). Thus, since \( L_I \) is non-negative,
\[
\mathbb{E}(L) \geq \frac{\lambda^2 \mathbb{E}(B^2)}{2(1 - \rho)} + \rho.
\] (31)
Combining (31) with (30) gives the desired result.

Another way to explain (31) is that the average number of customers in a queue with vacations is at least the average number of customers in a queue without vacations, a standard M/G/1 queue. \( \square \)

In order to investigate how tight the bounds derived in Theorem 4.1 are, we consider the case of exponentially distributed service times with mean 1. Further assume the vacation discipline of Example 3.2 with the vacation time distribution identical to the service time distribution. By Theorem 3.6 we then find for \( \psi(i) = 1/(i + 1) \) that \( \mathbb{E}(L) = 2\rho/(1 - \rho) \), while Theorem 4.1 gives \( \mathbb{E}(L) \geq \rho/(1 - \rho) \). So in this case the bound is off by a factor 2. We performed several numerical experiments for other de-activation probabilities \( \psi(\cdot) \). Two typical results are given in Figs. 1 and 2, which show simulation results for the average number of customers in the system for \( \psi(i) = 0.8^i \) and \( \psi(i) = (1/(i + 1))^{0.8} \), respectively. We further added the two lower bounds derived in Theorem 4.1, the inverse function bound (30) and the bound that follows from the Fuhrmann–Cooper decomposition (31). Note that we used a log-lin scale for graphical reasons.

For \( \psi(i) = 0.8^i \) we see that the simulation results are close to (31) for values of \( \rho \) close to 1, i.e., the bound in Theorem 4.1 seems rather tight in heavy traffic and the average number of customers is close to the average number of customers in a standard M/G/1 queue.
For $\psi(i) = (1/(i+1))^{0.8}$ the simulation results are close to the inverse function bound (30) for values of $\rho$ close to 1, i.e., the bound in Theorem 4.1 seems rather tight in heavy traffic for this choice for $\psi(\cdot)$ as well. In Sect. 5 we will prove that the bound in Theorem 4.1 is asymptotically exact in heavy traffic for the cases considered here, so that the order-of-magnitude of the stationary queue length is $O(1/(1-\rho))$ or $O(\psi^{-1}(1-\rho))$.

The next lemma presents a stochastic comparison result for two processes $\{L(t), \sigma(t)\}_{t\geq 0}$ and $\{\hat{L}(t), \hat{\sigma}(t)\}_{t\geq 0}$ with de-activation probabilities $\psi(\cdot)$ and $\hat{\psi}(\cdot)$, respectively.

**Lemma 4.2** For the vacation discipline described in Example 3.2, and assuming that $\hat{\psi}(i) \geq \psi(i)$, $i \geq 0$, $\hat{L}(0) = L(0)$ and $\hat{\sigma}(0) = \sigma(0) = 0$, $\{\hat{L}(t)\}_{t\geq 0} \overset{st}{\geq} \{L(t)\}_{t\geq 0}$.

**Proof** This can be proved using a coupling $\{(L^*(t), \sigma^*(t))\}_{t\geq 0}, \{(\hat{L}^*(t), \hat{\sigma}^*(t))\}_{t\geq 0}$ between $\{L(t), \sigma(t)\}_{t\geq 0}$ and $\{\hat{L}(t), \hat{\sigma}(t)\}_{t\geq 0}$. That is, we can construct the sample
path of the coupled system recursively such that, marginally, this sample path obeys the same probabilistic laws as the original process and we can prove that $\hat{L}^*(t) \geq L^*(t)$ for all $t \geq 0$.

A detailed proof of this lemma can be found in [4] and requires a careful analysis of all possible scenarios that can occur. We will only give a sketch of the main ideas here as the complete proof is very tedious and not particularly insightful.

In the sample path construction we make sure that (i) arrivals in both systems happen at the same time; (ii) the $n$th service takes the same amount of time in both systems; (iii) the $n$th ‘wait time’, i.e., the time $V$ described in Example 3.2, is equal in both systems; (iv) the system with de-activation probability $\hat{\psi}(\cdot)$ always de-activates if the system with de-activation probability $\psi(\cdot)$ de-activates, if the total numbers of customers in both systems are equal and a service ends in both systems.

It is easily seen that it is possible to construct sample paths such that the marginal paths obey the same probabilistic laws as the original processes in this way. Further, to see that $\hat{L}^*(t) \geq L^*(t)$ for all $t \geq 0$, first note that, by assumption, $\hat{L}^*(0) = L^*(0)$ and the sample paths are identical for the first period of time until the system with $\hat{\psi}(\cdot)$ de-activates, while the system with $\psi(\cdot)$ stays active. After that the system with $\hat{\psi}(\cdot)$ trails the system with $\psi(\cdot)$ until, possibly, at some time epoch the total time spent serving customers is equal again in both systems.

The first time epoch, say $t^*$ at which the latter can happen is, by (ii) and (iii), always a time epoch at which a ‘wait time’ ends in the system with $\psi(\cdot)$ and a service is completed in the system with $\hat{\psi}(\cdot)$ (these events happen at the exact same time by construction). If the system with $\hat{\psi}(\cdot)$ does not de-active at $t^*$, or if the systems are empty, we see that the sample paths will be identical again for some period of time. Further, if the systems are not empty and the system with $\hat{\psi}(\cdot)$ de-activates at $t^*$, we see that the system with $\hat{\psi}(\cdot)$ immediately starts trailing the system with $\psi(\cdot)$ again. In any case, following the same reasoning as above, it follows that $\hat{L}^*(t) \geq L^*(t)$ for all $t \geq 0$.

Note that we only proved the result of Lemma 4.2 for the vacation discipline of Example 3.2. For general vacation disciplines, which may depend on the arrival process, Lemma 4.2 does not always hold. It is for example not hard to see that Lemma 4.2 does not hold for $G_{X_0}(r) = r^{100}$, $G_{X}(r) = r$, $\psi(i) = 0$ for $i \geq 1$ and $\hat{\psi}(i) = 0.1$ for $i \geq 1$.

Combining Lemma 4.2 and Theorem 4.1 leads to a lower bound for the mean number of customers in a system with de-activation probability $\hat{\psi}(\cdot)$ if there exists a strictly decreasing convex function $\psi(\cdot)$ such that $\psi(i) \leq \hat{\psi}(i)$ for all $i$. We get

$$\mathbb{E}\{\hat{L}\} \geq \mathbb{E}\{L\} \geq \max \left\{ \psi^{-1}\left(\frac{1 - \rho}{\mathbb{E}\{X\}}\right), \frac{\lambda^2\mathbb{E}\{B^2\}}{2(1 - \rho)} + \rho \right\}.$$  

4.2 State-dependent vacation lengths

In this subsection we consider the vacation disciplines obeying Scenario 2. For this class of vacation disciplines we find the following bounds.
Theorem 4.3 Consider Scenario 2 and \( \psi(i) = 1, i \geq 0 \).

(i) If \( f(\cdot) \) is a strictly increasing, unbounded and concave function, then

\[
\mathbb{E}\{L\} \geq \frac{\lambda^2 \mathbb{E}\{B^2\}}{2(1 - \rho)} + \rho + f^{-1}\left(\frac{\lambda}{1 - \rho}\right).
\] (32)

(ii) If \( f(\cdot) \) is a strictly increasing convex function, then

\[
\mathbb{E}\{L\} \leq \frac{\lambda^2 \mathbb{E}\{B^2\}}{2(1 - \rho)} + \rho + f^{-1}\left(\frac{\lambda}{1 - \rho}\right).
\] (33)

Proof In steady state, the mean number of activations per unit of time equals the mean number of de-activations per unit of time, i.e.,

\[
\mathbb{P}\{\sigma = 1\} \cdot \frac{1}{\mathbb{E}\{B\}} = \mathbb{E}\{f(L_I)\} \mathbb{P}\{\sigma = 0\},
\] (34)

where \( L_I \) denotes the number of customers during a non-serving (vacation) period.

If \( f(\cdot) \) is concave, it follows by Jensen’s inequality that

\[
\mathbb{E}\{f(L_I)\} \leq f\left(\mathbb{E}\{L_I\}\right).
\]

Since \( f(\cdot) \) is increasing, we thus get, as \( \mathbb{P}\{\sigma = 1\} = \rho \) and \( \mathbb{P}\{\sigma = 0\} = 1 - \rho \),

\[
\mathbb{E}\{L_I\} \geq f^{-1}\left(\frac{\lambda}{1 - \rho}\right).
\]

The Fuhrmann–Cooper decomposition (7) implies

\[
\mathbb{E}\{L\} = \frac{\lambda^2 \mathbb{E}\{B^2\}}{2(1 - \rho)} + \rho + \mathbb{E}\{L_I\},
\]

yielding (32).

The bound in Eq. (33) follows by symmetry. \( \square \)

Note that \( f(i) = vi \) is both convex and concave. We thus find an exact result for this activation function,

\[
\mathbb{E}\{L\} = \frac{2\lambda + v\lambda^2 \mathbb{E}\{B^2\}}{2v(1 - \rho)} + \rho.
\]

This also follows from the generating function, which we derived in Theorem 3.4.

In order to investigate how tight the bounds in Theorem 4.3 are for activation functions other than \( f(i) = vi \), we performed several numerical experiments. Two typical results are displayed in Figs. 3 and 4, showing for exponentially distributed service times with mean one the average number of customers in the system for
Fig. 3 Average number of customers in the system for \( f(i) = 1.25^i - 1 \) and \( \rho \in [0, 1) \)

Fig. 4 Average number of customers in the system for \( f(i) = i^{0.8} \) and \( \rho \in [0, 1) \)

\( f(i) = 1.25^i - 1 \) and \( f(i) = i^{0.8} \), respectively. For these activation functions we see that the simulated results are relatively close to their corresponding bounds for all values of \( \rho \). In Sect. 5 we will prove that the bounds in Theorem 4.3 are in fact asymptotically sharp in heavy traffic.

The next lemma presents a stochastic comparison result for two processes \( \{L(t), \sigma(t)\}_{t \geq 0} \) and \( \{\hat{L}(t), \hat{\sigma}(t)\}_{t \geq 0} \) with activation rates \( f(\cdot) \) and \( \hat{f}(\cdot) \), respectively.

**Lemma 4.4** For Scenario 2, and assuming that \( \hat{f}(i) \leq f(i), \psi(i) = 1, i \geq 0, \hat{L}(0) = L(0) \) and \( \hat{\sigma}(0) = \sigma(0) = 0, \{\hat{L}(t)\}_{t \geq 0} \geq_{st} \{L(t)\}_{t \geq 0} \).

**Proof** The proof of this lemma can be found in [4]. It proceeds along similar lines as the proof of Lemma 4.2, i.e., it is based on a coupling \( \{L^*(t), \sigma^*(t)\}_{t \geq 0}, \{\hat{L}^*(t), \hat{\sigma}^*(t)\}_{t \geq 0} \) between \( \{L(t), \sigma(t)\}_{t \geq 0} \) and \( \{\hat{L}(t), \hat{\sigma}(t)\}_{t \geq 0} \).
In this sample path construction we make sure that (i) arrivals in both systems happen at the same time; (ii) the $n$th service takes the same amount of time in both systems; (iii) the system with activation rate $f(\cdot)$ always activates if the system with activation rate $\hat{f}(\cdot)$ activates, if the total numbers of customers in both systems are equal and both systems are inactive.

Following a similar reasoning as in Lemma 4.2 it can be verified that $\hat{L}^*(t) \geq L^*(t)$ for all $t \geq 0$ for this sample path construction.

Combining Lemma 4.4 and Theorem 4.3 leads to an upper bound for the mean number of customers in a system with activation rate $\hat{f}(\cdot)$ if there exists a strictly increasing convex function $f(\cdot)$ such that $f(i) \leq \hat{f}(i)$ for all $i$. We get

$$\mathbb{E}[\hat{L}] \leq \mathbb{E}[L] \leq \frac{\lambda^2 \mathbb{E}[B^2]}{2(1 - \rho)} + \rho + f^{-1}\left(\frac{\lambda}{1 - \rho}\right).$$

Similarly, we find a lower bound for the mean number of customers in a system if there exists a strictly increasing unbounded concave function $f(\cdot)$ such that $f(i) \geq \hat{f}(i)$ for all $i$.

5 Heavy-traffic results

In this section we study the heavy-traffic behavior of the system. In particular, we derive the stationary distribution of the scaled number of customers in the system in heavy traffic, $L/E\{L\}$ for $\rho \uparrow 1$. More precisely, we let $\lambda$ vary and study the system when $\lambda$ approaches $1/E\{B\}$.

As an important side result, we obtain the limiting distribution of the stationary scaled sojourn time as $\rho \uparrow 1$ as well. For this we consider the Laplace–Stieltjes transform of $S/E\{S\}, \mathbb{E}[e^{-wS/E\{S\}}]$, with $w \geq 0$. By virtue of the distributional form of Little’s law [16], the PASTA property and a level crossings argument we know that

$$G_L(r) = \mathbb{E}\{e^{-\lambda(1-r)S}\},$$

or equivalently, for $\mathbb{E}\{L\} \geq w$,

$$\mathbb{E}\{e^{-wS/E\{S\}}\} = G_L(1 - \frac{1}{\mathbb{E}\{L\}}w) = \mathbb{E}\{(1 - \frac{1}{\mathbb{E}\{L\}}w)^{\mathbb{E}\{L\}}\}^{\frac{1}{\mathbb{E}\{L\}}}. $$

Noting that $\mathbb{E}\{L\} \rightarrow \infty$ as $\rho \uparrow 1$ and using a generalized version of the continuous mapping theorem, see e.g., [15], we then find as $\rho \uparrow 1$

$$S/E\{S\} \overset{d}{\rightarrow} W \text{ if and only if } L/E\{L\} \overset{d}{\rightarrow} W.$$

Here $W$ denotes some non-negative random variable and $\overset{d}{\rightarrow}$ denotes convergence in distribution.

Springer
Note that $L$ and $X_i$ in general depend on the value of $\rho$, which is not fixed in this section. To emphasize this we will therefore write $G_L(r, \rho)$ for the generating function of $L$ and $G_{X_i}(r, \rho)$ for the generating function of $X_i$ in this section. Similarly, we write $K(r, \rho)$ and $Y(r, \rho)$ for $K(\cdot)$ and $Y(\cdot)$ as defined in (10) and (11). Further, in order to analyze the system in heavy traffic, we need to make some technical assumptions on the vacation distribution in heavy traffic. That is, in this section we will consider Scenario 3 as described below and Scenario 2 with functions $f(\cdot)$ that grow monotonically to infinity.

**Scenario 3** $X_i \overset{d}{=} X_{\geq 0}$ for all $i \geq 1$ and either $X_0 \overset{d}{=} X$ and $\mathbb{P}\{X = 0\} = 0$ or $X_0 \overset{\text{st}}{=} X$ and $\mathbb{P}\{X = 0\} > 0$. Further, $\mathbb{E}\{X_i^+\} = \lim_{\rho \uparrow 1} \mathbb{E}\{X_i\}$ exists and is finite for all $i \geq 0$, and

$$\lim_{\rho \uparrow 1} \left( \frac{\partial}{\partial \rho} G_{X_i}(r, \rho) \right)_{r = e^{-(1-\rho)u}} = 0,$$

for all $i \geq 0$ and $u \geq 0$.

The additional assumptions in Scenario 3, compared to Scenario 1, ensure that the vacation discipline behaves nicely when $\lambda$ approaches $1/\mathbb{E}\{B\}$, i.e., when $\rho \uparrow 1$ the vacation distribution for $\rho$ is similar to the vacation distribution for $\rho - \epsilon$ for small $\epsilon$, which is a desirable property from both a practical and theoretical perspective.

One example of a vacation discipline that belongs to Scenario 3 is the prototypical example of the back-off mechanism used in wireless networks described in Example 3.2. For $i \geq 1$ we get

$$\frac{\partial}{\partial \rho} G_{X_i}(r, \rho) = \frac{1 - r}{\mathbb{E}\{B\}} \tilde{V}'((1 - r)\rho/\mathbb{E}\{B\}),$$

and thus (35) holds because $\mathbb{E}\{V\} < \infty$. For $i = 0$ it can be checked in a similar way that (35) holds.

Denote by $\exp(\beta)$ a random variable having an exponential distribution with mean $1/\beta$, and denote by $\Gamma(\alpha, \beta)$ a random variable having a gamma distribution with shape parameter $\alpha$ and rate parameter $\beta$. Define $R_B = \mathbb{E}\{B^2\}/(2\mathbb{E}\{B\})$, the mean residual service time, and $\nu_B = R_B/\mathbb{E}\{B\}$.

**Theorem 5.1** Consider Scenario 3 and $\psi(i) = a^i$ with $0 \leq a < 1$, $i \geq 0$. Then,

$$(1 - \rho)L \overset{d}{\rightarrow} \exp(\nu_B^{-1})$$

as $\rho \uparrow 1$.

**Proof** Consider the Laplace–Stieltjes transform of $(1 - \rho)L$, with $u \geq 0$, and note that

$$\mathbb{E}\{e^{-(1-\rho)uL}\} = G_L(e^{-(1-\rho)u}, \rho).$$

We will now use Theorem 3.2 to prove (36). For this define $h(\rho) = e^{-(1-\rho)u}$ and note that then,
\[
\frac{Y(h(\rho), \rho)}{Y(1, \rho)} = \frac{(1 - \rho) \tilde{B}(\rho(1 - h(\rho))/\mathbb{E}(B))(1 - G_X(h(\rho), \rho))}{\mathbb{E}(X)(B(\rho(1 - h(\rho))/\mathbb{E}(B)) - h(\rho))}.
\] (38)

Applying l’Hôpital’s rule twice,
\[
\lim_{\rho \uparrow 1} \frac{Y(h(\rho), \rho)}{Y(1, \rho)} = \frac{1}{1 + \nu_B h'(1)} \left(1 + \lim_{\rho \uparrow 1} \left(\frac{1}{h'(\rho)\mathbb{E}(X)} \frac{\partial}{\partial \rho} G_X(r, \rho) \bigg|_{r=h(\rho)}\right)\right).
\] (39)

By continuity of \(Y(\cdot, \cdot)\),
\[
\lim_{\rho \uparrow 1} Y(a^i h(\rho), \rho) = Y(a^i, 1),
\]
for \(i \geq 1\). From Theorem 3.2 we then find for \(X_0 \overset{d}{=} X\) that
\[
\lim_{\rho \uparrow 1} G_L(h(\rho), \rho) = \frac{1}{1 + \nu_B h'(1)} = \frac{1}{1 + \nu_B u},
\]
which, by Lévy’s continuity theorem, gives (36) in case \(X_0 \overset{d}{=} X\).

We also find
\[
\lim_{\rho \uparrow 1} \left(\frac{G_X(h(\rho), \rho) - G_{X_0}(h(\rho), \rho)}{1 - G_X(h(\rho), \rho)}\right) = \lim_{\rho \uparrow 1} \left(\frac{\frac{\partial}{\partial \rho} G_X(r, \rho) + h'(\rho)\mathbb{E}(X) - \frac{\partial}{\partial \rho} G_{X_0}(r, \rho) - h'(\rho)\mathbb{E}(X)}{\mathbb{E}(X)} \bigg|_{r=h(\rho)}\right)
\]
\[
= \frac{\mathbb{E}(X_0) - \mathbb{E}(X)}{\mathbb{E}(X)} = \frac{K(1, \rho)}{Y(1, \rho)},
\]
and, for \(i \geq 1\),
\[
\lim_{\rho \uparrow 1} \frac{G_X(a^i h(\rho), \rho) - G_{X_0}(a^i h(\rho), \rho)}{1 - G_X(a^i h(\rho), \rho)} = \frac{G_X(a^i, 1) - G_{X_0}(a^i, 1)}{1 - G_X(a^i, 1)}.
\]

Further, from Theorem 3.2 we find, for \(X_0 >_{st} X\),
\[
G_L(h(\rho), \rho) = \frac{Y(h(\rho), \rho)}{Y(1, \rho)} \sum_{j=0}^{\infty} \frac{G_X(a^j h(\rho), \rho) - G_{X_0}(a^j h(\rho), \rho)}{1 - G_X(a^j h(\rho), \rho)} \prod_{i=1}^{j} Y(a^i h(\rho), \rho) \sum_{j=0}^{\infty} \frac{G_X(a^j, \rho) - G_{X_0}(a^j, \rho)}{1 - G_X(a^j, \rho)} \prod_{i=1}^{j} Y(a^i, \rho),
\]
as
\[
K(r, \rho) = Y(r, \rho) \frac{G_X(r, \rho) - G_{X_0}(r, \rho)}{1 - G_X(r, \rho)}.
\]
We thus obtain
\[
\lim_{\rho \uparrow 1} G_L(h(\rho), \rho) = \frac{1}{1 + h'(1)u_B} = \frac{1}{1 + u u_B},
\]
which proves (36).

It is striking that the result in Theorem 5.1 is independent of the precise assumption on when the server returns from a vacation. In fact, the behavior is similar to the heavy-traffic behavior of a standard M/G/1 queue without vacations [17].

Remember that in this paper we assume \( E\{B^2\} < \infty \). This assumption is needed in the proof of Theorem 5.1, but not in the proof of Theorem 3.2. If the service time distribution has a tail behavior like \( t^{-k} \) with \( 1 < k \leq 2 \), i.e., the service time has finite mean and infinite variance, we can prove along similar lines as the proof of Theorem 5.1 that then the heavy-traffic behavior is similar to that of a standard M/G/1 queue without vacations as well [6].

If the server de-activates less frequently than in Theorem 5.1, then one would expect the same result as in Theorem 5.1. The next theorem proves this result for the vacation discipline of Example 3.2. Furthermore, we will prove a similar result for vacation disciplines in Scenario 2 with an aggressive activation function \( f(\cdot) \).

**Theorem 5.2** For the vacation discipline described in Example 3.2 and \( \psi(i) \leq a^i \) with \( a \in [0, 1) \), \( i \geq 0 \), and for Scenario 2 with \( \psi(i) = 1, i \geq 0 \), and \( f(\cdot) \) a strictly increasing continuous and convex function with \( \lim_{i \to \infty} i^{-1} f^{-1}(i) = 0 \),
\[
(1 - \rho)L \xrightarrow{d} \exp(u_B^{-1}) \text{ as } \rho \uparrow 1.
\] (40)

**Proof** First assume the vacation discipline of Example 3.2 is used with \( \psi(i) \leq a^i, a \in [0, 1) \). By Lemma 4.2 we have \( L \leq_{st} L_{a^i} \), where \( L_{a^i} \) denotes a random variable with the steady-state distribution of the number of customers in the system with \( \psi(i) = a^i \) for all \( i \). Further, \( L \geq_{st} L_{M/G/1} \). The result now follows from Theorem 5.1.

Now assume Scenario 2 with \( \psi(i) = 1, i \geq 0 \), and \( f(\cdot) \) a strictly increasing continuous and convex function with \( \lim_{i \to \infty} i^{-1} f^{-1}(i) = 0 \). From (33) we get, because \( \lim_{i \to \infty} i^{-1} f^{-1}(i) = 0 \),
\[
\lim_{\rho \uparrow 1} (1 - \rho) E\{L\} \leq u_B.
\] (41)

Now consider the random variable \( W = (1 - \rho)(L - L_{M/G/1}) \) and note that \( W \) is nonnegative because \( L \geq_{st} L_{M/G/1} \) and \( \rho < 1 \). Therefore,
\[
E\{|W|\} = E\{W\} = E\{(1 - \rho)L\} - E\{(1 - \rho)L_{M/G/1}\}.
\]

As
\[
\lim_{\rho \uparrow 1} (1 - \rho) E\{L_{M/G/1}\} = u_B.
\]
we find from (41) that \( \mathbb{E}[|W|] = 0 \), hence \( W \) converges in mean to 0. Using Slutsky’s theorem we then get

\[
(1 - \rho)L = W + (1 - \rho)L_{M/G/1} \xrightarrow{d} \exp(v_B^{-1}) \text{ as } \rho \uparrow 1,
\]

which completes the proof. \( \square \)

Even though the results of Theorem 5.2 only hold in heavy traffic when \( \rho \uparrow 1 \), the regime that we are interested in, they can be useful for more moderate values of \( \rho \) for some cases. Consider for example the case \( \psi(i) = a^i \), exponentially distributed service times with mean 1 and the vacation discipline of Example 3.2 with the vacation time distribution identical to the service time distribution. Define the relative error as

\[
\Delta = \frac{\hat{L}_{HT}}{L_{sim}} - 1, \quad (42)
\]

with \( \hat{L}_{HT} = 1/(1 - \rho) \) the expected number of customers in the system using the heavy-traffic approximation and \( L_{sim} \) the average number of customers in the system in a simulation. Table 1 presents \( \Delta \) for various values of \( a \) and \( \rho \). From these results we see that the heavy-traffic approximation works well for large values of \( \rho \) when \( a \) is small enough. For larger values of \( a \) this is not the case and the heavy-traffic results should then be used with care.

Next we consider vacation disciplines that are less aggressive. First we will consider Scenario 3 with \( \psi(i) \) inversely proportional to the queue length and Scenario 2 with a linear activation rate \( f(i) \). For these cases we will show that the heavy-traffic behavior does depend on the vacation scenario.

**Theorem 5.3** Consider Scenario 3 and \( \psi(i) = 1/(i + 1), \ i \geq 0 \). Then,

\[
(1 - \rho)L \xrightarrow{d} \Gamma(1 + \mathbb{E}\{X^*\}v_B^{-1}, v_B^{-1}) \text{ as } \rho \uparrow 1. \quad (43)
\]

Similarly, for Scenario 2 with \( \psi(i) = 1 \) and \( f(i) = vi, \ i \geq 0 \),

\[
(1 - \rho)L \xrightarrow{d} \Gamma(1 + 1/(vR_B), v_B^{-1}) \text{ as } \rho \uparrow 1. \quad (44)
\]

| Table 1 | Relative error for \( \psi(i) = a^i \) for different values of \( a \) and \( \rho \) |
|---------|----------------|----------------|----------------|----------------|
| \( a \) \( \times \) \( \rho \) | 0.8 | 0.9 | 0.95 | 0.99 |
| 0.25 | 0.0012 | -0.0209 | -0.0120 | 0.0020 |
| 0.35 | -0.0317 | -0.0323 | -0.0287 | 0.0049 |
| 0.45 | -0.0774 | -0.0592 | -0.0510 | 0.0378 |
| 0.55 | -0.1465 | -0.1148 | -0.0645 | 0.0200 |
| 0.65 | -0.2367 | -0.1821 | -0.1227 | -0.0117 |
| 0.75 | -0.3703 | -0.3051 | -0.2051 | -0.0732 |
| 0.85 | -0.5576 | -0.4877 | -0.3740 | -0.1400 |
| 0.95 | -0.8306 | -0.7929 | -0.7118 | -0.3534 |

\( \square \) Springer
Proof The proof for Scenario 3 proceeds along similar lines as the proof of Theorem 5.1. We consider the Laplace–Stieltjes transform of $(1 - \rho)L, \mathbb{E}[e^{-(1-\rho)uL}]$, with $u \geq 0$, and use (37) and Theorem 3.3 to prove (43). For this define $h(\rho) = e^{-(1-\rho)u}$ and note that,

$$
\lim_{\rho \uparrow 1} \int_{h(\rho)}^{1} \frac{Y(x, \rho)}{x} \, dx = \lim_{\rho \uparrow 1} \int_{h(\rho)}^{1-h(\rho)} \frac{1}{x} \left( \frac{1}{1 + \rho^2 u B s} + O(1 - \rho) \right) \, ds
$$

Using Taylor expansion and noting that $s = O(1 - \rho)$ as $\rho \uparrow 1$ in the integration domain,

$$
\int_{h(\rho)}^{1} \frac{Y(x, \rho)}{x} \, dx = \frac{\mathbb{E}[X^*]}{1 - \rho} \left( \frac{1 - \rho}{\rho^2 u B} \log \left( 1 + \frac{\rho^2 u B (1 - h(\rho))}{1 - \rho} \right) + O((1 - \rho)^2) \right).
$$

Thus,

$$
\lim_{\rho \uparrow 1} \exp \left( - \int_{h(\rho)}^{1} \frac{Y(x, \rho)}{x} \, dx \right) = \left( 1 + u u B \right)^{-\mathbb{E}[X^*]/u B}.
$$

We now find for $X_0 \overset{d}{=} X$, using (39) and Theorem 3.3, that

$$
\lim_{\rho \uparrow 1} GL(h(\rho), \rho) = \left( \frac{1}{1 + u u B} \right)^{1+\mathbb{E}[X^*]/u B}, \tag{45}
$$

which, by Lévy’s continuity theorem, gives (43) in case $X_0 \overset{d}{=} X$.

For $X_0 \succ X$, using (13) and (22),

$$
P\{L = 0\}K(h(\rho), \rho) = \frac{GL_{M/G/1}(h(\rho), \rho) \left( \mathbb{E}[X_0]G_{X_0}^{res}(h(\rho), \rho) - \mathbb{E}[X]G_{X}^{res}(h(\rho), \rho) \right)}{\mathbb{E}[X_0] - \mathbb{E}[X] + \mathbb{E}[X] \frac{1}{1-\rho} \int C(x, \rho) \, dx}, \tag{46}
$$

with

$$
C(x, \rho) = GL_{M/G/1}(x) \left( \mathbb{E}[X_0]G_{X_0}^{res}(x, \rho) - \mathbb{E}[X]G_{X}^{res}(x, \rho) \right)e^{\alpha(x, \rho)}.
$$

Thus $\lim_{\rho \uparrow 1} P\{L = 0\}K(h(\rho), \rho) = 0$, as $C(x, \rho) > 0$. 

Springer
Similarly,
\[
\lim_{\rho \uparrow 1} \frac{1}{1 - \rho} \frac{\int_0^{h(\rho)} C(x, \rho) \, dx}{\mathbb{E}[X_0] - \mathbb{E}[X] + \frac{\mathbb{E}[X]}{1 - \rho} \int_0^{h(\rho)} C(x, \rho) \, dx} = \frac{1}{\mathbb{E}[X]}.
\]

Hence by Theorem 3.3 and (45) we find for \(X_0 >_\text{st} X\) that
\[
\lim_{\rho \uparrow 1} G_L(h(\rho), \rho) = \left( \frac{1}{1 + u\nu_B} \right)^{1 + u\nu_B} \mathbb{E}[X^*]
\]
as well, which proves (43).

For Scenario 2 with \(\psi(i) = 1\) and \(f(i) = \nu i, i \geq 0\), note that
\[
\int_{h(\rho)}^1 \frac{-\lambda(1 - x)}{\nu(B(\lambda(1 - x)) - x)} \, dx = -\frac{\lambda}{\nu} \int_0^{1-h(\rho)} \frac{s}{\widetilde{B}(\lambda s) - 1 + s} \, ds.
\]

Using Taylor expansion gives
\[
\int_{h(\rho)}^1 \frac{-\lambda(1 - x)}{\nu(B(\lambda(1 - x)) - x)} \, dx = -\frac{\lambda}{\nu} \int_0^{1-h(\rho)} \left( \frac{s}{-\rho s + \rho^2 \nu_B s^2 + s} + O(1) \right) \, ds
\]
\[
= -\frac{\rho}{\nu(1 - \rho)\mathbb{E}[B]} \int_0^{1-h(\rho)} \left( \frac{1}{1 + \frac{\rho^2 \nu_B}{1 - \rho}} + O(1 - \rho) \right) \, ds
\]
\[
= -\frac{\rho}{\nu(1 - \rho)\mathbb{E}[B]} \left( 1 - \frac{1}{\rho^2 \nu_B} \log \left( 1 + \frac{\rho^2 \nu_B (1 - h(\rho))}{1 - \rho} \right) + O((1 - \rho)^2) \right).
\]

Thus,
\[
\lim_{\rho \uparrow 1} \int_{h(\rho)}^1 \exp \left( \frac{-\lambda(1 - x)}{\nu(B(\lambda(1 - x)) - x)} \right) \, dx = \left( 1 + u\nu_B \right)^{-1/(vR_B)},
\]

which, using Theorem 3.4 and Lévy’s continuity theorem, gives (44).

\[\square\]

Instead of using the result in Theorem 5.1 one could also use the differential Eqs. (23) and (24) to prove Theorem 5.3 directly.
Note that, for any value of $\rho$, the result in Theorem 5.3 is in fact exact for the case considered in Theorem 3.6 and exact in the first moment for the case considered in Theorem 3.5.

We thus see that the heavy-traffic behavior of $L$ does depend on the specific parameters if the vacation disciplines of Theorem 5.3 are used. We further see that the number of customers still scales like $1 - \rho$ in heavy traffic, i.e., $(1 - \rho)L$ converges to a random variable. This is not the case anymore for vacation disciplines that are even less aggressive as the next theorem states.

**Theorem 5.4** For the vacation discipline described in Example 3.2 and $\psi(i) = 1/(i + 1)^{\alpha}$ with $\alpha \in (0, 1)$, $i \geq 0$, and for Scenario 2 with $\psi(i) = 1$ and $f(i) = \nu^{i^{\alpha}}$ with $\alpha \in (0, 1)$, $i \geq 0$,

$$\frac{L}{\mathbb{E}\{L\}} \overset{d}{\to} 1 \text{ as } \rho \uparrow 1. \quad (47)$$

In particular, for the vacation discipline described in Example 3.2, with the vacation probability inversely proportional to the queue length raised to the power $\alpha$, $\alpha \in (0, 1)$, and for Scenario 2 with a linear activation rate raised to the power $\alpha$, $\alpha \in (0, 1)$, $i \geq 0$,

$$\lim_{\rho \uparrow 1} (1 - \rho)^{1/\alpha} \mathbb{E}\{L\} = \mathbb{E}\{X^*\}^{1/\alpha},$$

and for Scenario 2,

$$\lim_{\rho \uparrow 1} (1 - \rho)^{1/\alpha} \mathbb{E}\{L\} = \mathbb{E}\{B\}^{-1/\alpha} \nu^{-1/\alpha}.$$  

We thus see for the vacation discipline described in Example 3.2 with the vacation probability inversely proportional to the queue length raised to the power $\alpha$, $\alpha \in (0, 1)$, and for Scenario 2 with a linear activation rate raised to the power $\alpha$, $\alpha \in (0, 1)$, that $(1 - \rho)L$ diverges. In fact, we see that for these vacation disciplines the number of customers in the system scales like $(1 - \rho)^{1/\alpha}$ in heavy traffic. We further see that, using this appropriate heavy-traffic scaling, the scaled number of users in the system in heavy traffic has a degenerate distribution.

In order to prove Theorem 5.4 we first introduce some additional notation. For any function $g(\cdot) : [0, \infty) \mapsto [0, \infty)$ define, for $a < 1$, $b > 1$ and $x \in [0, \infty)$,

$$\gamma_{a, b}(x) = \frac{(b - 1)g(ax) + (1 - a)g(bx)}{(b - a)g(x)}.$$  

Further define

$$\kappa_{a, b} = 1 - \sup_x \gamma_{a, b}(x),$$

and

$$\chi_{a, b} = 1 - \inf_x \gamma_{a, b}(x).$$

The proof of Theorem 5.4 is based on the following proposition.
Proposition 5.5 Assume $g(\cdot)$ is concave and $\kappa_{a,b} > 0$ for any $a < 1$ and $b > 1$, or $g(\cdot)$ is convex and $\chi_{a,b} < 0$ for any $a < 1$ and $b > 1$. If

$$\lim_{\rho \uparrow 1} \frac{\mathbb{E}\{g(W)\}}{g(\mathbb{E}\{W\})} = 1$$

then

$$\frac{W}{\mathbb{E}\{W\}} \xrightarrow{d} 1 \text{ as } \rho \uparrow 1.$$

It is possible that a proof of Proposition 5.5 is available in the literature, but for completeness a self-contained proof is provided in the Appendix.

Having established Proposition 5.5, we can now prove Theorem 5.4.

Proof (of Theorem 5.4) For the vacation discipline described in Example 3.2 and $\psi(i) = 1/(i + 1)^\alpha$ with $\alpha \in (0, 1), i \geq 0$, we know from Theorem 4.1 that

$$\mathbb{E}\{L\} \geq \psi^{-1}\left(\frac{1 - \rho}{\mathbb{E}\{X\}}\right),$$

or, as $\psi^{-1}(i) = i^{-1/\alpha} - 1$,

$$\lim_{\rho \uparrow 1} (1 - \rho)^{1/\alpha} \mathbb{E}\{X\}^{-1/\alpha} \mathbb{E}\{L\} \geq 1.$$

Now consider the system with $\hat{\psi}(i) = 1$ for $i \leq \lceil \psi^{-1}(\beta) \rceil$ and $\hat{\psi}(i) = \beta$ for $i > \lceil \psi^{-1}(\beta) \rceil$, where $\beta > 0$ and, for stability, $\beta < (1 - \rho)/\mathbb{E}\{X\}$. Thus, by construction, $\hat{\psi}(i) \geq \psi(i)$ for all $i$. Further assume that this system uses a vacation discipline similar to that of Example 3.2, but with a slight modification; the server only activates if at least $\lceil \psi^{-1}(\beta) \rceil + 1$ customers are present in the system, instead of at least 1. That is, we have $\lceil \psi^{-1}(\beta) \rceil$ permanent customers. It follows immediately from Lemma 4.2 that $\hat{L} \geq_{st} L$. Further, using (8) we find

$$G_{\hat{L}}(r) = \frac{\mathbb{P}\{\hat{L} = \lceil \psi^{-1}(\beta) \rceil\} \tilde{B}(\lambda(1 - r)) (G_X(r) - G_{X_0}(r)) r^{\lceil \psi^{-1}(\beta) \rceil}}{\tilde{B}(\lambda(1 - r)) - r - (1 - G_X(r)) \beta},$$

with

$$\mathbb{P}\{\hat{L} = \lceil \psi^{-1}(\beta) \rceil\} = \frac{(1 - \rho - \beta \mathbb{E}\{V\})(1 - \tilde{V}(\lambda))}{\lambda \mathbb{E}\{V\} \tilde{V}(\lambda)}.$$

Now take $\beta = \frac{(1 - \rho)(1 - \delta)}{\mathbb{E}\{X\}}, \delta > 0$, and note that from (49)

$$\mathbb{E}\{\hat{L}\} = \lceil \psi^{-1}(\frac{(1 - \rho)(1 - \delta)}{\mathbb{E}\{X\}}) \rceil + C(\rho),$$
with \( \lim_{\rho \uparrow 1} C(\rho)(1 - \rho) < \infty \). Using that \( \psi^{-1}(i) = i^{-1/\alpha} - 1 \), we get

\[
(1 - \rho)^{1/\alpha} \mathbb{E}\{X\}^{-1/\alpha} \mathbb{E}\{\hat{L}\} \leq (1 - \delta)^{-1/\alpha} + (1 - \rho)^{1/\alpha} \mathbb{E}\{X\}^{-1/\alpha} C(\rho),
\]

and, hence, as \( 0 < \alpha < 1 \),

\[
\lim_{\rho \uparrow 1} (1 - \rho)^{1/\alpha} \mathbb{E}\{X\}^{-1/\alpha} \mathbb{E}\{\hat{L}\} \leq \lim_{\rho \uparrow 1} (1 - \delta)^{-1/\alpha} + (1 - \rho)^{1/\alpha} \mathbb{E}\{X\}^{-1/\alpha} C(\rho) \leq (1 - \delta)^{-1/\alpha},
\]

for any \( \delta > 0 \). Thus, as \( \hat{L} \geq_{st} L \), we find

\[
\lim_{\rho \uparrow 1} (1 - \rho)^{1/\alpha} \mathbb{E}\{X\}^{-1/\alpha} \mathbb{E}\{L\} \leq 1.
\]

Therefore, using Eq. (48),

\[
\lim_{\rho \uparrow 1} (1 - \rho)^{1/\alpha} \mathbb{E}\{X\}^{-1/\alpha} \mathbb{E}\{L\} = 1,
\]

or

\[
\lim_{\rho \uparrow 1} \frac{\psi(\mathbb{E}\{L\})}{1 - \rho} = \frac{1}{\mathbb{E}\{X\}}.
\]

From (29) we find

\[
\mathbb{E}\{\psi(L)\} = \frac{1}{\mathbb{E}\{X\}} (1 - \rho - \mathbb{P}\{L = 0\}),
\]

and hence, because \( \mathbb{P}\{L = 0\}/(1 - \rho) \rightarrow 0 \) as \( \rho \uparrow 1 \) by using Lemma 4.2 and an argument similar to (46),

\[
\lim_{\rho \uparrow 1} \frac{\mathbb{E}\{\psi(L)\}}{1 - \rho} = \frac{1}{\mathbb{E}\{X\}}.
\]

Combining this with (51) we find

\[
\lim_{\rho \uparrow 1} \frac{\mathbb{E}\{\psi(L)\}}{\psi(\mathbb{E}\{L\})} = 1.
\]

Further, because \( \psi(\cdot) \) is strictly convex,

\[
\gamma_{a,b}(x) = \frac{(b - 1)(1 + ax)^{-\alpha} + (1 - a)(1 + bx)^{-\alpha}}{(b - a)(1 + x)^{-\alpha}}
\]

\[
= \frac{b - 1}{b - a} \frac{(1 + ax)^{-\alpha}}{1 + x} + \frac{1 - a}{b - a} \frac{(1 + bx)^{-\alpha}}{1 + x}
\]

\[
> \left( \frac{b - 1}{b - a} \frac{1 + ax}{1 + x} + \frac{1 - a}{b - a} \frac{1 + bx}{1 + x} \right)^{-\alpha} = 1,
\]

\[\text{Springer}\]
and the statement for the vacation discipline described in Example 3.2 follows from Proposition 5.5.

The proof for Scenario 2 with \( \psi(i) = 1 \) and \( f(\cdot) = \nu i^\alpha \) with \( \alpha \in (0, 1) \), \( i \geq 0 \), proceeds along similar lines. First, using Theorem 4.3 we know

\[
\mathbb{E}(L) \geq \frac{\lambda^2 \mathbb{E}(B^2)}{2(1-\rho)} + \rho + f^{-1}\left(\frac{\lambda}{1-\rho}\right),
\]

or, as \( f^{-1}(i) = (i/\nu)^{1/\alpha} \),

\[
\lim_{\rho \uparrow 1}(1-\rho)^{1/\alpha} \mathbb{E}(B)^{1/\alpha} v^{1/\alpha} \mathbb{E}(L) \geq 1.
\] (52)

Now consider the system with \( \hat{f}(i) = 0 \) for \( i \leq \lceil f^{-1}(\beta) \rceil \) and \( \hat{f}(i) = \beta \) for \( i > \lceil f^{-1}(\beta) \rceil \), where \( \beta > 0 \) and, for stability, \( \beta > \frac{\rho}{(1-\rho)\mathbb{E}(B)} \). Thus, by construction, \( \hat{f}(i) \leq f(i) \) for all \( i \geq 0 \). Further take \( \hat{\psi}(i) = 1 \) for all \( i \geq 0 \). We know from Lemma 4.4 that \( \hat{L} \geq_{st} L \).

We can find the generating function of \( \hat{L} \) using (8). We can also find this generating function by noting that this system behaves as an M/G/1 queue with \( \lceil f^{-1}(g) \rceil \) permanent customers and service requirement \( B + \exp(\beta) \), i.e., the time required to serve a customer is the sum of the vacation time and the service time.

Now take \( \beta = \frac{\rho}{(1-\rho)(1-\delta)\mathbb{E}(B)} \), \( \delta > 0 \), which gives

\[
\mathbb{E}(\hat{L}) = \lceil f^{-1}(\frac{\rho}{(1-\rho)(1-\delta)\mathbb{E}(B)}) \rceil + C(\rho),
\]

with \( \lim_{\rho \uparrow 1} C(\rho)(1-\rho) < \infty \). We then find in a similar way as before that

\[
\lim_{\rho \uparrow 1}(1-\rho)^{1/\alpha} \mathbb{E}(B)^{1/\alpha} v^{1/\alpha} \mathbb{E}(L) \leq 1,
\]

and hence, using (52),

\[
\lim_{\rho \uparrow 1}(1-\rho)^{1/\alpha} \mathbb{E}(B)^{1/\alpha} v^{1/\alpha} \mathbb{E}(L) = 1.
\]

Noting that \( L_{M/G/1}/\mathbb{E}(L) \xrightarrow{d} 0 \) as \( \rho \uparrow 1 \) we thus find, using the Fuhrmann–Cooper decomposition (7),

\[
\lim_{\rho \uparrow 1}(1-\rho) f(\mathbb{E}(L)) = \frac{1}{\mathbb{E}(B)}.
\]

Further, from (34) we find

\[
\lim_{\rho \uparrow 1}(1-\rho)\mathbb{E}(f(L)) = \frac{1}{\mathbb{E}(B)},
\]
so that

\[
\lim_{\rho \uparrow 1} \frac{\mathbb{E}\{f(L_I)\}}{\psi(\mathbb{E}\{L_I\})} = 1. \tag{53}
\]

Finally, because \(f(\cdot)\) is strictly concave,

\[
\gamma_{a,b}(x) = \frac{(b - 1)a^\alpha + (1 - a)b^\alpha}{b - a} < \left(\frac{b - 1}{b - a}a + \frac{1 - a}{b - a}b\right)^\alpha = 1,
\]

and we find (47) by invoking (53) and Proposition 5.5. \(\square\)

The proof of Theorem 5.4 can be simplified if we assume \(\mathbb{E}\{B^3\} < \infty\) and \(\mathbb{E}\{X_i^3\} < \infty\) for all \(i \geq 0\). In that case we can find \(\mathbb{E}\{L^2\}\) along similar lines as we found \(\mathbb{E}\{L\}\) in the proof of Theorem 5.4. It then follows that \(\lim_{\rho \uparrow 1} \frac{\mathbb{E}\{L^2\}}{\mathbb{E}\{L\}^2} = 1\) in this case, so that the assertion in Theorem 5.4 follows from Chebyshev’s inequality.

Even though the results in Theorem 5.4 were derived to demonstrate the trichotomy in heavy traffic, it is interesting to see when the results can be used as an approximation for systems that are not in heavy traffic. For this consider exponentially distributed service times with mean 1 and the vacation discipline of Example 3.2 with the vacation time distribution identical to the service time distribution. The relative error (42) for \(\psi(i) = 1/(i + 1)^\alpha\) is given in Table 2 for different values of \(\alpha\) and \(\rho\). From these results we again see that the heavy-traffic approximation works well for large values of \(\rho\) when the activation scheme is ‘far away’ from the intermediate case considered in Theorem 5.3, i.e., when \(\alpha\) is small enough. The accuracy degrades when the activation scheme is ‘closer’ to the intermediate case, i.e., when \(\alpha\) is close to 1.

### 6 Conclusions

In this paper we have obtained results for queues with random back-offs. Such random back-offs can be modeled by rates of activating during a back-off period \(f(\cdot)\) and by the probability of initiating a back-off period after a service completion \(\psi(\cdot)\). For various choices of \(f(\cdot)\) and \(\psi(\cdot)\), and under some additional assumptions, we have obtained exact expressions for the distribution of the number of customers in the system in Sect. 3 and bounds for the mean stationary number of customers in the

| \(\alpha\) \(\backslash\) \(\rho\) | 0.8 | 0.9 | 0.95 | 0.99 |
|---|---|---|---|---|
| 0.35 | -0.1378 | -0.0058 | 0.0073 | 0.0068 |
| 0.45 | -0.2388 | -0.0883 | -0.0430 | 0.0088 |
| 0.55 | -0.3153 | -0.1986 | -0.1160 | 0.0274 |
| 0.65 | -0.3801 | -0.2778 | -0.1941 | -0.0601 |
| 0.75 | -0.4269 | -0.3557 | -0.3113 | -0.1563 |
| 0.85 | -0.4596 | -0.4241 | -0.3924 | -0.3253 |
| 0.95 | -0.4863 | -0.4795 | -0.4640 | -0.4416 |
system in Sect. 4. These results were employed to derive heavy-traffic limit theorems in Sect. 5, which showed the existence of a clear trichotomy, that can be best explained through the function \( \psi(\cdot) \), taking the vacation time independent of the queue length. Clearly, in order for the system to be stable when \( \rho \uparrow 1 \), \( \psi(\cdot) \) should eventually, as the number of customers increases, go to zero. This condition is also sufficient; see Lemma 3.1. Roughly speaking (for details and further assumptions see Sect. 5), the queueing system with back-offs can display three modes of operation, depending on the asymptotic decay rate of \( \psi(\cdot) \). These three modes can be understood as follows:

(i) The case \( \psi(i) = 1/(i+1) \) represents a balanced regime, in which the heavy-traffic behavior is influenced by both the back-off probability \( \psi(\cdot) \) and the heavy-traffic behavior of the corresponding system without back-offs. Hence, large queue sizes typically build up according to sample paths that exhibit exceptional (interrupted) busy periods and exceptional sequences of back-off periods. The number of customers \( L \) is of the order \( O((1-\rho)^{-1}) \), and the more detailed information in Theorem 5.3 reveals a gamma distribution containing information of the arrival process, service times, and the back-off function.

(ii) When \( \psi(\cdot) \) decays faster than \( 1/(i+1) \), for instance \( \psi(i) = O(a^i) \) with \( a \in (0,1) \), it is shown that the heavy-traffic behavior of the system is as if the back-off periods do not exist. The intuition is that when \( \rho \uparrow 1 \), the system spends most of the time in states of large queue sizes in which the probability of initiating a back-off becomes negligible. Indeed, in Theorem 5.1 it is shown that for \( \psi(i) = a^i \) with \( a \in (0,1) \) the heavy-traffic behavior of the system is the same as that of an M/G/1 system without back-offs.

(iii) When \( \psi(\cdot) \) decays slower than \( 1/(i+1) \), for instance for \( \psi(i) = O((i+1)^{-a}) \) with \( a \in (0,1) \), it is shown that the heavy-traffic behavior of the system is completely determined by the back-offs. Theorem 5.4 says that for \( \psi(i) = 1/(i+1)^a \) with \( a \in (0,1) \) the mean number of customers is \( O((1-\rho)^{-1/a}) \), while the stationary distribution of the number of customers is degenerate and thus strongly concentrated around its mean. Hence, for systems with such back-offs, the heavy-traffic behavior is entirely different from that of systems without back-off (the M/G/1 system in this case).

Another relevant observation that follows from the analysis is that the order of magnitude of the number of customers in the system \( L \) in heavy traffic is independent of the vacation and service time distribution in all three modes.

The revealed trichotomy for the single-node system provides some important insights for the wireless networks equipped with back-off rules similar to those discussed in Sect. 1, because the single-node system provides a best-case scenario for networks with multiple nodes. To see this, consider a network that is in heavy traffic because the aggregated traffic intensity in some clique, a set of nodes of which at most one can be active at the same time, tends to 1. The total number of customers in this clique behaves like the number of customers in the corresponding single-node system with two modifications. First, the probability to go into back-off is based on a subset of all customers. Hence the network will be in back-off more often if \( \psi(\cdot) \) is decreasing and the total number of customers in both systems were equal. Second, the length of the vacation period is the minimum over the back-off lengths of all non-blocked nodes,
which might change during the vacation period. We thus see that the vacation length is at least equal to the minimum back-off length of a node in the clique assuming none of the nodes is prevented from activating. Hence, taking this minimum as the actual vacation length in the corresponding single-node system, we see that vacations in the network always take at least as long as in the single-node system. As both modifications intuitively have a negative impact on the delay performance, it seems reasonable to assume that the total number of customers in the network is at best equal to the total number of customers in the corresponding single-node system.

We thus see that more aggressive activation schemes can potentially improve the delay performance. On the other hand, however, these aggressive activation schemes may fail to achieve maximum stability and hence can be unstable in heavy traffic. The results in [10, 13, 21, 24] only guarantee maximum stability for general networks when $\psi(i) = O(1/(\log(i) + 1))$, which, based on the analysis in this paper, might result in very poor delay performance in heavy traffic. An interesting topic for further research is to establish for which scenarios the delay performance in the network is roughly equal to the delay performance of the corresponding single-node system.

Acknowledgments This work was supported by Microsoft Research through its Ph.D. Scholarship Programme, an ERC starting Grant and a TOP Grant from NWO. We thank J.A.C. Resing for bringing the work of Sevast’yanov to our attention.

Appendix: Preliminary results and proofs

This appendix contains a few technical lemmas and some proofs that have been relegated from the main text. To make this appendix self-contained we restate some results from the main text.

Lemma 7.1 (i) If $X_0 \overset{d}{=} X$, then,

$$\prod_{i=0}^{\infty} Y(a^i r),$$

with $0 \leq a < 1$, converges for all $r \in [0, 1]$.

(ii) If $X_0 \overset{\text{st}}{>} X$, then,

$$\sum_{j=0}^{\infty} K(a^j r) \prod_{i=0}^{j-1} Y(a^i r),$$

with $0 \leq a < 1$, converges for all $r \in [0, 1]$.

Proof To prove case (i) first note that this infinite product converges if and only if

$$\sum_{i=0}^{\infty} (Y(a^i r) - 1)$$
converges. To prove convergence of this infinite series we will use the ratio test (d’Alembert’s criterion). We have, with \( h(r) = \tilde{B}(\lambda(1 - r)) \) and \( k(r) = \tilde{B}(\lambda(1 - r))G_X(r) \),

\[
\lim_{i \to \infty} \left| \frac{Y(a^{i+1}r) - 1}{Y(a^i r) - 1} \right| = \lim_{i \to \infty} \frac{(-a^i r + h(a^i r))(a^{i+1}r - k(a^{i+1}r))}{(-a^{i+1}r + h(a^{i+1}r))(a^i r - k(a^i r))} = \lim_{i \to \infty} \frac{a^{i+1}r - k(a^{i+1}r)}{a^i r - k(a^i r)}.
\]

By l’Hôpital’s rule,

\[
\lim_{i \to \infty} \frac{a^{i+1}r - k(a^{i+1}r)}{a^i r - k(a^i r)} = \lim_{i \to \infty} \frac{1 + \lambda G_X(a^{i+1}r)\tilde{B}'(\lambda(1 - a^{i+1}r)) - \tilde{B}(\lambda(1 - a^{i+1}r))G_X'(a^{i+1}r)}{1 + \lambda G_X(a^i r)\tilde{B}'(\lambda(1 - a^i r)) - \tilde{B}(\lambda(1 - a^i r))G_X'(a^i r)}.
\]

We thus find

\[
\lim_{i \to \infty} \left| \frac{Y(a^{i+1}r) - 1}{Y(a^i r) - 1} \right| = a < 1,
\]

proving case (i).

For case (ii) note that

\[
\lim_{n \to \infty} \frac{K(a^{n+1}r)\prod_{i=0}^{n} Y(a^i r)}{K(a^n r)\prod_{i=0}^{n-1} Y(a^i r)} = Y(0) < 1,
\]

for all \( r \) as \( 0 \leq a < 1 \). Thus, by the ratio test, the series in case (ii) converges.

\[\square\]

**Lemma 7.2** If \( \alpha y + (1 - \alpha)z = \alpha' y' + (1 - \alpha')z' \), with \( 0 \leq \alpha, \alpha' \leq 1 \) and \( y' \leq y \leq z' \), then

(i) If \( g(\cdot) \) is a concave function,

\[
\alpha g(y) + (1 - \alpha)g(z) \geq \alpha' g(y') + (1 - \alpha')g(z').
\]  \hfill (54)

(ii) If \( g(\cdot) \) is a convex function,

\[
\alpha g(y) + (1 - \alpha)g(z) \leq \alpha' g(y') + (1 - \alpha')g(z').
\]  \hfill (55)

**Proof** Since \( y' \leq y \leq z \leq z' \), there exist \( 0 \leq \alpha_y, \alpha_z \leq 1 \), such that \( y = \alpha_y y' + (1 - \alpha_y)z' \), and \( z = \alpha_z y' + (1 - \alpha_z)z' \). It follows from the equality \( \alpha y + (1 - \alpha)z = \alpha' y' + (1 - \alpha')z' \) that \( \alpha' = \alpha \alpha_y + (1 - \alpha)\alpha_z \), and \( 1 - \alpha' = \alpha(1 - \alpha_y) + (1 - \alpha)(1 - \alpha_z) \).

Further, if \( g(\cdot) \) is concave,

\[
\alpha_y g(y') + (1 - \alpha_y)g(z') \leq g(y),
\]
and

\[ \alpha z g(y') + (1 - \alpha z)g(z') \leq g(z). \]

We may then write

\begin{align*}
\alpha g(y) + (1 - \alpha)g(z) &\geq \alpha[\alpha_y g(y') + (1 - \alpha_y)g(z')] + (1 - \alpha)[\alpha_z g(y') + (1 - \alpha_z)g(z')] \\
&= [\alpha\alpha_y + (1 - \alpha)\alpha_z]g(y') + [\alpha(1 - \alpha_y) + (1 - \alpha)(1 - \alpha_z)]g(z') \\
&= \alpha'g(y') + (1 - \alpha')g(z'),
\end{align*}

which completes the proof for case (i). The inequality in (55) follows by symmetry. \( \Box \)

**Corollary 7.3** For all \( x \), if \( a' \leq a < 1, b' \geq b > 1 \), then

(i) If \( g(\cdot) \) is a concave function, \( \gamma_{a',b'}(x) \leq \gamma_{a,b}(x) \leq 1 \) and thus \( \kappa_{a',b'} \geq \kappa_{a,b} \geq 0 \).

(ii) If \( g(\cdot) \) is a convex function, \( \gamma_{a',b'}(x) \geq \gamma_{a,b}(x) \geq 1 \) and thus \( \chi_{a',b'} \leq \chi_{a,b} \leq 0 \).

**Proof** Taking \( y = ax, y' = a'x, z = bx, z' = b'x, \alpha = (b - 1)/(b - a) \), and \( \alpha' = (b' - 1)/(b' - a') \) in (54), we obtain for \( g(\cdot) \) concave,

\[ \frac{(b - 1)g(ax) + (1 - a)g(bx)}{b - a} = \alpha g(y) + (1 - \alpha)g(z) \geq \alpha' g(y') + (1 - \alpha')g(z') \]

\[ = \frac{(b' - 1)g(a'x) + (1 - a')g(b'x)}{b' - a'}, \]

which yields the statement for concave \( g(\cdot) \).

The assertion for convex \( g(\cdot) \) follows by symmetry. \( \Box \)

Let \( W \) henceforth be a nonnegative integer-valued random variable with probability distribution \( p(x) = \mathbb{P}\{W = x\} \). For any \( y \geq 0 \), define \( F(y) = \mathbb{P}\{W \leq y\} = \mathbb{P}\{W \leq \lfloor y \rfloor\} \), with pseudo inverse

\[ F^{-1}(u) = \inf\{y : F(y) \geq u\} \]

for any \( u \in [0, 1] \), so that we may write

\[ \mathbb{E}\{g(W)\} = \int_{u=0}^{1} g(F^{-1}(u))du, \]

and in particular

\[ \mathbb{E}\{W\} = \int_{u=0}^{1} F^{-1}(u)du. \]
For compactness, denote $\hat{F}^{-1}(u) = F^{-1}(u)/\mathbb{E}[W]$, 

$$x_1(\epsilon_1) = \frac{1}{\epsilon_1} \int_{u=0}^{\epsilon_1} \hat{F}^{-1}(u) du,$$

and

$$x_2(\epsilon_2) = \frac{1}{\epsilon_2} \int_{u=1-\epsilon_2}^{1} \hat{F}^{-1}(u) du.$$

**Lemma 7.4** Let $0 < \epsilon_1 \leq F(\mathbb{E}[W])$, $0 < \epsilon_2 \leq 1 - F(\mathbb{E}[W])$, so that $x_1(\epsilon_1) \leq \hat{F}^{-1}(\epsilon_1) \leq 1$ and $x_2(\epsilon_2) \geq \hat{F}^{-1}(1-\epsilon_2) \geq 1$, with

$$\epsilon_1 x_1(\epsilon_1) + \epsilon_2 x_2(\epsilon_2) = \epsilon_1 + \epsilon_2,$$

or equivalently,

$$\int_{u=\epsilon_1}^{1-\epsilon_2} \hat{F}^{-1}(u) du = 1 - \epsilon_1 - \epsilon_2.$$

(i) If $g(\cdot)$ is a concave function,

$$(\epsilon_1 + \epsilon_2)x_1(\epsilon_1)x_2(\epsilon_2) \leq 1 - \frac{\mathbb{E}[g(W)]}{g(\mathbb{E}[W])}. \quad (56)$$

(ii) If $g(\cdot)$ is a convex function,

$$(\epsilon_1 + \epsilon_2)x_1(\epsilon_1)x_2(\epsilon_2) \geq 1 - \frac{\mathbb{E}[g(W)]}{g(\mathbb{E}[W])}. \quad (57)$$

**Proof** Write

$$\mathbb{E}[g(W)] = \int_{u=0}^{\epsilon_1} g(F^{-1}(u)) du + \int_{u=\epsilon_1}^{1-\epsilon_2} g(F^{-1}(u)) du + \int_{u=1-\epsilon_2}^{1} g(F^{-1}(u)) du. \quad (58)$$

Because of Jensen’s inequality we find for concave $g(\cdot)$

$$\int_{u=\epsilon_1}^{1-\epsilon_2} g(F^{-1}(u)) du \leq (1 - \epsilon_1 - \epsilon_2)g\left(\frac{1}{1 - \epsilon_1 - \epsilon_2} \int_{u=\epsilon_1}^{1-\epsilon_2} F^{-1}(u) du\right)$$

$$= (1 - \epsilon_1 - \epsilon_2)g(\mathbb{E}[W]).$$
Invoking Jensen’s inequality once again,

\[
\int_{u=0}^{\epsilon_1} g(F^{-1}(u))du + \int_{u=1-\epsilon_2}^{1} g(F^{-1}(u))du \\
\leq \epsilon_1 g \left( \frac{1}{\epsilon_1} \int_{u=0}^{\epsilon_1} F^{-1}(u)du \right) + \epsilon_2 g \left( \frac{1}{\epsilon_2} \int_{u=1-\epsilon_2}^{1} F^{-1}(u)du \right) \\
= \epsilon_1 g(x_1(\epsilon_1)E[W]) + \epsilon_2 g(x_2(\epsilon_2)E[W]) \\
= y_{x_1(\epsilon_1),x_2(\epsilon_2)}(E[W])(\epsilon_1 + \epsilon_2)g(E[W]) \\
\leq (1 - \kappa_{x_1(\epsilon_1),x_2(\epsilon_2)})(\epsilon_1 + \epsilon_2)g(E[W]).
\]

Substituting the above two inequalities in (58) we obtain the statement of the lemma for concave \( g(\cdot) \). The assertion for convex \( g(\cdot) \) follows from symmetry. \( \square \)

**Lemma 7.5** Let \( 0 < \epsilon_1 \leq F(E[W]) \), \( 0 < \epsilon_2 \leq 1 - F(E[W]) \), so that \( x_1(\epsilon_1) \leq \hat{F}^{-1}(\epsilon_1) \leq 1 \) and \( x_2(\epsilon_2) \geq \hat{F}^{-1}(1 - \epsilon_2) \geq 1 \), with

\[
\epsilon_1 x_1(\epsilon_1) + \epsilon_2 x_2(\epsilon_2) = \epsilon_1 + \epsilon_2,
\]

or equivalently,

\[
\int_{u=\epsilon_1}^{1-\epsilon_2} \hat{F}^{-1}(u)du = 1 - \epsilon_1 - \epsilon_2.
\]

(i) If \( g(\cdot) \) is a concave function,

\[
\kappa_{x_1(\epsilon_1),x_2(\epsilon_2)} \geq \max\{\kappa_{\hat{F}^{-1}(\epsilon_1),1+\frac{\epsilon_1}{\epsilon_2}(1-\hat{F}^{-1}(\epsilon_1))}, \kappa_{1-\frac{\epsilon_2}{\epsilon_1}(\hat{F}^{-1}(1-\epsilon_2)-1),\hat{F}^{-1}(1-\epsilon_2)}\}.
\]

(ii) If \( g(\cdot) \) is a convex function,

\[
\chi_{x_1(\epsilon_1),x_2(\epsilon_2)} \leq \min\{\chi_{\hat{F}^{-1}(\epsilon_1),1+\frac{\epsilon_1}{\epsilon_2}(1-\hat{F}^{-1}(\epsilon_1))}, \chi_{1-\frac{\epsilon_2}{\epsilon_1}(\hat{F}^{-1}(1-\epsilon_2)-1),\hat{F}^{-1}(1-\epsilon_2)}\}.
\]

**Proof** Observing that

\[
x_2(\epsilon_2) \geq \hat{F}^{-1}(1 - \epsilon_2),
\]

we obtain

\[
\epsilon_1 x_1(\epsilon_1) \leq \epsilon_1 + \epsilon_2 - \epsilon_2 \hat{F}^{-1}(1 - \epsilon_2) = \epsilon_1 + \epsilon_2(1 - \hat{F}^{-1}(1 - \epsilon_2)).
\]
In addition,

$$x_1(\epsilon_1) \leq \hat{F}^{-1}(\epsilon_1),$$

yielding

$$x_1(\epsilon_1) \leq \min\{\hat{F}^{-1}(\epsilon_1), 1 - \frac{\epsilon_2}{\epsilon_1}(\hat{F}^{-1}(1 - \epsilon_2) - 1)\}.$$ 

Likewise,

$$x_2(\epsilon_2) \geq \max\{\hat{F}^{-1}(1 - \epsilon_2), 1 + \frac{\epsilon_1}{\epsilon_2}(1 - \hat{F}^{-1}(\epsilon_1))\}.$$ 

Combining the above two inequalities and using Corollary 7.3 completes the proof. 

\[\square\]

**Proposition 5.5** Assume $g(\cdot)$ is concave and $\kappa_{a,b} > 0$ for any $a < 1$ and $b > 1$, or $g(\cdot)$ is convex and $\chi_{a,b} < 0$ for any $a < 1$ and $b > 1$. If

$$\lim_{\rho \uparrow 1} \frac{\mathbb{E}\{g(W)\}}{g(\mathbb{E}\{W\})} = 1,$$

then

$$\frac{W}{\mathbb{E}\{W\}} \xrightarrow{d} 1 \text{ as } \rho \uparrow 1.$$ 

**Proof** Take $\delta > 0$ and $\epsilon_1 = F((1 - \delta)\mathbb{E}\{W\})$. Then either $\epsilon_1 = 0$, or $0 < \epsilon_1 \leq F(\mathbb{E}\{W\})$ and $x_1(\epsilon) \leq \hat{F}^{-1}(\epsilon_1) \leq 1 - \delta$. In the latter case, define $\epsilon_2^* = 1 - \hat{F}^{-1}(\mathbb{E}\{W\})$, and observe that

$$\int_{u=\epsilon_1}^{1-\epsilon_2^*} \hat{F}^{-1}(u)du \leq 1 - \epsilon_1 - \epsilon_2^*,$$

while

$$\int_{u=\epsilon_1}^{1} \hat{F}^{-1}(u)du \geq 1 - \epsilon_1.$$ 

Hence, by continuity, there must exist an $\epsilon_2 \in (0, \epsilon_2^*)$ with $x_2(\epsilon_2) > 1$ and

$$\int_{u=\epsilon_1}^{1-\epsilon_2} \hat{F}^{-1}(u)du = 1 - \epsilon_1 - \epsilon_2,$$
so that the assumptions of Lemmas 7.4 and 7.5 are satisfied. Applying these two lemmas then yields for concave $g(\cdot)$

$$
\kappa \hat{F}^{-1}(\epsilon_1), 1 + \epsilon_1 (1 - \hat{F}^{-1}(\epsilon_1)) \leq \kappa \hat{F}^{-1}(\epsilon_1), 1 + \epsilon_1 \left(1 - \frac{\epsilon_1}{\epsilon_2} (1 - \hat{F}^{-1}(\epsilon_1))\right) \rightarrow 0 \text{ as } \rho \uparrow 1.
$$

This means that $\epsilon_1 = \mathbb{P}[W \leq (1 - \delta)\mathbb{E}[W]] \rightarrow 0$ as $\rho \uparrow 1$. A similar argument shows that $\mathbb{P}[W \geq (1 + \delta)\mathbb{E}[W]] \rightarrow 0$ as $\rho \uparrow 1$. It now follows from the definition of convergence in probability that $\mathbb{E}[W] \rightarrow 1$ as $\rho \uparrow 1$ if $g(\cdot)$ is concave.

The proof for convex $g(\cdot)$ follows by symmetry. \qed

References

1. Borst, S.C.: Polling systems with multiple coupled servers. Queueing Syst. 20, 369–393 (1995)
2. Bouman, N., Borst, S.C., van Leeuwaarden, J.S.H.: Delays and mixing times in random-access networks. SIGMETRICS Perf. Eval. Rev. 41(1), 117–128 (2013). Proceedings of SIGMETRICS 2013 Conference
3. Bouman, N., Borst, S.C., van Leeuwaarden, J.S.H.: Delay performance of backlog-based random access. SIGMETRICS Perf. Eval. Rev. 39(2), 32–34 (2011). Proceedings of Performance 2011 Conference
4. Bouman, N., Borst, S.C., Boxma, O.J., van Leeuwaarden, J.S.H.: Queues with random back-offs. Technical Report arXiv:1302.3144 (2013)
5. Bouman, N., Borst, S.C., van Leeuwaarden, J.S.H., Proutière, A.: Backlog-based random access in wireless networks: Fluid limits and delay issues. In Proceedings of ITCS 23, 2011
6. Boxma, O.J., Cohen, J.W.: Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions. Queueing Syst. 33, 177–204 (1999)
7. Capiński, M., Kopp, E.: Measure, Integral and Probability. Springer, London (2004)
8. Crabill, T.B.: Sufficient conditions for positive recurrence of specially structured Markov chains. Oper. Res. 16, 858–867 (1968)
9. Fuhrmann, S.W., Cooper, R.B.: Stochastic decompositions for the M/G/1 queue with generalized vacations. Oper. Res. 33, 1117–1129 (1985)
10. Ghaderi, J., Srikant, R.: On the design of efficient CSMA algorithms for wireless networks. In Proceedings of CDC 2010 Conference, 2010
11. Ghaderi, J., Borst, S.C., Whiting P.A.: Backlog-based random-access in wireless networks: Fluid limits and instability issues. In Proceedings of WiOpt 2012 Conference, 2012
12. Harris, C.M., Marchal, W.G.: State dependence in M/G/1 server vacation models. Oper. Res. 36(4), 560–565 (1988)
13. Jiang, L., Shah, D., Shin, J., Walrand, J.: Distributed random access algorithm: Scheduling and congestion control. IEEE Trans. Inf. Theory 56(12), 6182–6207 (2010)
14. Jiang, L., Leconte, M., Ni, J., Srikant, R., Walrand J.: Fast mixing of parallel Glauber dynamics and low-delay CSMA scheduling. In Proceedings of Infocom 2011 Mini-Conference, 2011
15. Kallenberg, O.: Foundations of Modern Probability. Springer, New York (1997)
16. Keilson, J., Servi, L.: A distributional form of Little’s law. Oper. Res. Lett. 7, 223–227 (1988)
17. Kingman, J.F.C.: On queues in heavy traffic. J. R. Stat. Soc. Ser. B 24(2), 383–392 (1962)
18. Lottiniezhad, M., Marbach, P.: Throughput-optimal random access with order-optimal delay. In Proceedings of Infocom 2011 Conference, 2011
19. Ni, J., Tan, B., Srikant, R.: Q-CSMA: queue length based CSMA/CA algorithms for achieving maximum throughput and low delay in wireless networks. In Proceedings Infocom 2010 Mini-Conference, 2010
20. Pakes, A.G.: Some conditions for ergodicity and recurrence of Markov chains. Oper. Res. 17, 1058–1061 (1969)
21. Rajagopalan, S., Shah, D., Shin, J.: Network adiabatic theorem: An efficient randomized protocol for contention resolution. In Proceedings of ACM SIGMETRICS/Performance 2009 Conference, 2009
22. Sevast’yanov, B.: Limit theorems for branching stochastic processes of special form. Th. Prob. Appl. 2, 321–331 (1957)
23. Shah, D., Shin J.: Delay-optimal queue-based CSMA. In Proceedings of ACM SIGMETRICS 2010 Conference, 2010
24. Shah, D., Shin, J., Tetali P.: Medium access using queues. In Proceedings of FOCS 2011 Conference, 2011
25. Takagi, H.: Queueing Analysis: Vacation and Priority Systems. North-Holland, Amsterdam (1991)