The Foreground-Background Queue
Nuijens, M.F.M.

Citation for published version (APA):
Nuijens, M. F. M. (2004). The Foreground-Background Queue

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
This chapter contains a number of different topics centred around the stationary queue length $Q$ in the M/G/1 FB queue. First we show that under a certain condition all moments of $Q$ are finite. We consider the mean of $Q$ in case of deterministic service times. Some heavy-traffic limits are established for the mean of $Q$. Finally, the cohort process in the stationary queue is described.

### 9.1 Moments of the stationary queue length

The main object of study in this chapter is the queue length in the stationary queue operating under the FB discipline. We start the chapter with a section devoted to the moments of the stationary queue length $Q$ in the M/G/1 FB queue. For the M/G/1 PS queue all moments of the stationary queue length $Q_{PS}$ exist since $P(Q_{PS} = n) = (1 - \rho)^{n}, n \geq 0$. In the M/G/1 FIFO queue the following holds for each $n \in \mathbb{N}$: $EQ_{FIFO}^{n} < \infty$ only if $EB^{n+1} < \infty$. This may be seen by using the Pollaczek-Khinchin transform for the queue length, see e.g. Kleinrock [31]. We shall
prove the following theorem for the moments of the stationary queue length in the FB queue.

**Theorem 9.1** Let $Q$ be the stationary queue length in the $M/G/1$ FB queue with generic service time $B$. If $EB^\alpha < \infty$ for some $\alpha > 1$, then all moments of $Q$ are finite.

Recall from Section 2.3 that the moment generating function $\phi(z)$ of the stationary queue length $Q$ satisfies

$$
\phi(z) = E_z Q = (1 - \rho) e^{\gamma(z)}, \quad \text{where} \quad \gamma(z) = -z \int_0^\infty \frac{\partial v(t, z)}{\partial z} dt. \quad (9.1)
$$

Here $v(t, z), t \geq 0$, is the unique nonnegative root of the equation (2.8). From the proof of relation (9.1) in Pechinkin [44], it follows that the function $v$ satisfies $v(t, z) = \lambda(1 - \psi_1(z))$, where $\psi_1(z)$ is the probability generating function of a random variable $Z_t$, which is such that $P(Z_t \in \{0, 1, 2, \ldots\}) = 1$. Hence

$$
v(t, z) = \lambda - \lambda \sum_{k=0}^\infty P(Z_t = k) z^k. \quad (9.2)
$$

Define $\gamma_n = \lim_{z \to 1} \gamma^{(n)}(z)$ for $n \in \mathbb{N}$. Then $\phi^{(1)}(z) = \gamma^{(1)}(z) \phi(z)$ and $\phi^{(1)}(1) = \gamma_1$ since $P(Q < \infty) = 1$. Likewise $\phi^{(2)}(1) = \gamma_2 + \gamma_1^2$ and in general $\phi^{(n)}(1)$ is a polynomial in $\gamma_1, \ldots, \gamma_n$ for every $n \in \mathbb{N}$. This leads us to the following lemma.

**Lemma 9.2** If $\gamma_j < \infty$ for all $1 \leq j \leq m$, then the $m$-th moment of $Q$ is finite.

**Proof** Fix an $m \in \mathbb{N}$. Then $\phi^{(m)}(1)$ is a linear polynomial in $EQ, \ldots, EQ^m$ in which the coefficient of $EQ^m$ is 1. Hence $EQ^m$ is a linear polynomial in $\phi^{(1)}(1), \ldots, \phi^{(m)}(1)$. But for every $j \in \mathbb{N}$ we may see $\phi^{(j)}(1)$ as a polynomial in $\gamma_1, \ldots, \gamma_j$. Hence $EQ^m$ is a polynomial in $\gamma_1, \ldots, \gamma_m$ and $EQ^m$ is finite if $\gamma_1, \ldots, \gamma_m$ are finite. \qed

**Lemma 9.3** Let $n \in \mathbb{N}$. Then $\gamma_n < \infty$ if

$$
a_k = \lim_{z \to 1} \frac{\partial^k}{\partial z^k} \int_0^\infty \frac{\partial v(t, z)}{\partial z} dt < \infty, \quad 1 \leq k \leq n. \quad (9.3)
$$

**Proof** From (9.2) it follows that the derivatives of $v(t, z)$ w.r.t. $z$ are negative and decreasing in $z$. Hence from (9.1) it follows that $\gamma_n$ is a polynomial in $a_1, \ldots, a_n$. \qed
To prove that (9.3) holds, we calculate the partial derivatives of \( v(t, z) \) w.r.t. \( z \). Differentiating both sides of equation (2.8) gives

\[
\frac{\partial v(t, z)}{\partial z} = \frac{\partial}{\partial z} \left( -\lambda \int_0^t e^{-v(t, z)x} dF(x) - \lambda z(1 - F(t))e^{-v(t, z)t} \right).
\]  

(9.4)

Since \( v(t, z) = \lambda - \lambda \psi_t(z) \), where \( \psi_t(z) \) is a probability generating function of a non-negative integer valued random variable, the integrand in (9.4) is bounded for \( 0 \leq z < 1 \). By a consequence of Lebesgue’s Dominated convergence theorem, we may interchange differentiation and integration in (9.4) and obtain

\[
\frac{\partial v(t, z)}{\partial z} = h(t, z) - \lambda(1 - F(t))e^{-v(t, z)t}.
\]  

(9.5)

where

\[
h(t, z) = \lambda \int_0^t x e^{-v(t, z)x} dF(x) + \lambda z(1 - F(t))te^{-v(t, z)t}.
\]  

(9.6)

Differentiation of (9.6) yields that for \( n = 1, 2, \ldots \)

\[
\frac{\partial^n h(t, z)}{\partial z^n} = \sum_{k=1}^n \lambda \left( \int_0^t x^{k+1} e^{-v(t, z)x} dF(x) + z(1 - F(t))t^{k+1}e^{-v(t, z)t} \right) P_{k,n} + (1 - F(t))t^k e^{-v(t, z)t} P_{k-1,n-1}.
\]  

(9.7)

where \( P_{k,n} \) is a homogeneous polynomial of degree \( k \) in \( \frac{\partial^j v(t, z)}{\partial z^j} \), \( j = 1, \ldots, n - k + 1 \), with coefficients in \( \mathbb{Z} \). So \( P_{k,n}(\alpha x_1, \ldots, \alpha x_{n-k+1}) = \alpha^k P_{k,n}(x_1, \ldots, x_{n-k+1}) \).

**Lemma 9.4** The partial derivatives of \( v(t, z) \) satisfy for \( t \geq 0 \) and \( 0 \leq z < 1 \),

\[
\frac{\partial^n v(t, z)}{\partial z^n} = \sum_{S_n} c(l, k_1, \ldots, k_m) \left( \frac{\lambda(1 - F(t))e^{-v(t, z)t}}{1 - h(t, z)} \right)^{l-1} \prod_{i=1}^m \frac{\partial^{k_i} h(t, z)}{\partial z^{k_i}},
\]  

(9.8)

where \( c(l, k_1, \ldots, k_m) \) \( \in \mathbb{Z} \) and

\[
S_n = \{(l, k_1, \ldots, k_m) : l \geq 1, 1 \leq k_1 \leq \ldots \leq k_m \}, \quad l + k_1 + \cdots + k_m = n \}.
\]

**Proof** By induction. For \( n = 1 \) (9.8) is equal to (9.5). Differentiating the RHS of (9.8) w.r.t. \( z \) yields the RHS of (9.8) for \( n + 1 \).

**Lemma 9.5** If \( EB^n < \infty \) for some \( \alpha > 1 \), then for every \( n \in \mathbb{N} \)

\[
\frac{\partial^n v(t, z)}{\partial z^n} \bigg|_{z=1} = O(t^{-\alpha}), \quad t \to \infty.
\]  

(9.9)
Chapter 9 The stationary queue length

**Proof** By induction. Assume that (9.9) holds for \( n \). Lemma 7.16 yields that \( 1 - F(t) = O(t^{-\alpha}) \) as \( t \to \infty \). Since \( v(t, 1) = 0 \), we have \( h(x, 1) = \lambda E(B \wedge x) \leq \rho < 1 \). Hence the first quotient in every summand in (9.8) is \( O(t^{-\alpha}) \). Proposition 7.16 yields that \( E(B \wedge t)^r = o(t^{-\alpha}) \) for all \( r > \alpha \). Applying this and the induction hypothesis to (9.7) yields

\[
\frac{\partial^k h(t, z)}{\partial z^k} \bigg|_{z=1} = O(t^{1-\alpha})
\]

for all \( k \leq n \). Hence every summand in (9.8) is \( O(t^{-\alpha}) \) and the proof is finished. □

**Proof of Theorem 9.1** Let \( n \in \mathbb{N} \) and define

\[
w_n(t, z) = \frac{\partial^n v(t, z)}{\partial z^n}.
\]

By (9.2), \( |w_n(t, z)| \) is well-defined for every \( t \) and increasing in \( 0 \leq z < 1 \). By the assumption that \( E B^\alpha < \infty \) for some \( \alpha > 1 \), Lemma 9.5 implies that \( |w_n(t, z)| \) is bounded by a function of order \( O(t^{-\alpha}) \) for \( t \to \infty \). Since \( w_n(t, z) \) is well-defined for every \( t \), it is integrable w.r.t. \( t \).

A corollary to Lebesgue’s Dominated convergence theorem then yields that in (9.3) we may first interchange integration and differentiation, and then the integral and the limit. By (9.2), the partial derivative \( \frac{\partial^n v(t, z)}{\partial z^n} \) is left continuous in \( z = 1 \). Using Lemma 9.5, we then obtain with the familiar abuse of notation.

\[
a_n = \lim_{z \to 1} \left| \frac{\partial^n}{\partial z^n} \int_0^\infty \frac{\partial v(t, z)}{\partial z} dt \right| = \left| \int_0^\infty \lim_{z \to 1} \left. \frac{\partial^{n+1} v(t, z)}{\partial z^{n+1}} \right|_{z=1} dt \right| = \left| \int_0^\infty O(t^{-\alpha}) dt \right| < \infty.
\]

By Lemma 9.3, it follows that \( \gamma_n \) is finite for all \( n \geq 1 \). Use Lemma 9.2 to complete the proof. □

The next proposition states that in the stationary queue, where \( E B < \infty \), \( EQ \) is finite. It is an open question whether higher moments are finite as well if only the first moment of \( B \) is assumed to be finite. We use the notation \( \rho(x) = \lambda E(B \wedge x) \).

**Proposition 9.6** In the stationary M/G/1 FB queue, \( EQ < \infty \).

**Proof** Let \( B_1 \) be an i.i.d. copy of \( B \). By (2.9), and since \( \rho(x) \leq \rho < 1 \).

\[
EQ = \lambda \int_0^\infty \left( \frac{\lambda E(B \wedge x)^2}{2(1-\rho(x))^2} + \frac{x}{1-\rho(x)} \right) F(x) dx
\]

\[
= E_{B_1} \lambda^2 E(B \wedge B_1)^2 \frac{1}{2(1-\rho(B_1))^2} + E_{B_1} \lambda B_1 \frac{1}{1-\rho(B_1)} < \infty \iff E(B \wedge B_1)^2 < \infty.
\]
9.2 The stationary M/D/1 queue

The distribution of \( B \land B_1 \) is given by

\[
P(B \land B_1 \leq x) = 1 - P(B \land B_1 > x) = 1 - P(B > x)^2 = 1 - (1 - F(x))^2.
\]

Since \( x(1 - F(x)) = o(1) \) by Lemma 7.16, we may bound \( x(1 - F(x)) \leq \gamma \) for some \( \gamma \in \mathbb{R} \). Hence

\[
E(B \land B_1)^2 = \int_0^\infty x^2 d(1 - (1 - F(x))^2) = -\int_0^\infty x^2 d(1 - F(x))^2
\]

\[
= -x^2(1 - F(x))^2|_{x=0}^{\infty} + 2 \int_0^\infty x(1 - F(x))^2 dx
\]

\[
\leq 2 \int_0^\infty \gamma(1 - F(x)) dx = 2\gamma EB.
\]

This completes the proof. \( \square \)

9.2 The stationary M/D/1 queue

In this section we discuss some properties of the stationary M/D/1 FB queue. First we show that deterministic service times maximise the expected queue length among all service-time distributions with the same mean. Then we compare the expected queue lengths in the stationary M/D/1 queues operating under FB and FIFO.

We begin this section by quoting the following lemma from Rai et al. [47].

**Lemma 9.7** The sojourn time \( V(x) \) in the stationary M/G/1 FB queue of a customer with service time \( x \) satisfies

\[
EV(x) \leq \frac{(2 - \rho)x}{2(1 - \rho)^2},
\]

and equality holds if and only if \( P(B = x) = 1 \).

**Proof** Theorem 2 in Rai et al. [47]. \( \square \)

**Corollary 9.8** Let \( Q \) be the stationary queue length in an M/G/1 FB queue with service-time distribution \( F \) and workload \( \rho \). Then

\[
EQ \leq \frac{\rho(2 - \rho)}{2(1 - \rho)^2}
\]

and equality holds if and only if the service-time distribution is a degenerate distribution, i.e. \( P(B = c) = 1 \) for some \( c > 0 \).
**Proof** Let $V$ be the sojourn time of a customer in an $M/G/1$ queue with service-time distribution $F$. Then by Little's law and Lemma 9.7,

$$EQ = \lambda EV = \lambda \int_0^\infty EV(x)dF(x) \leq \lambda \int \frac{(2-\rho)x}{2(1-\rho)^2}dF(x) = \frac{\rho(2-\rho)}{2(1-\rho)^2}.$$  \hspace{1cm} (9.11)

From Lemma 9.7 it follows that in (9.11) equality holds if and only if $F$ is degenerate. \hfill \Box

The inequality (9.10) is stated in Rai *et al.* [47] and the expression on the RHS of (9.10) for the expected queue length in the $M/D/1$ FB queue is mentioned in Yashkov [69]. The combination of these two seems to be new.

From Corollary 9.8 it follows that given $EB$ and $\lambda$, the $M/D/1$ queue maximises the mean queue length among all $M/G/1$ FB queues. This is not entirely unexpected in the light of an observation made before: the FB discipline behaves poorly for deterministic service times, since all customers in a busy period have to wait till the end of the busy period to leave the queue. Furthermore, the degenerate distribution has certainly a decreasing mean residual life. Hence Theorem 2.2 yields that for the $M/D/1$ queue the FB discipline has the largest value of $EQ$ among all disciplines in $\mathcal{D}$. Yet, this extremity property is of a different nature than the extremity property shown in Corollary 9.8. Theorem 2.2 compares queues with different disciplines, whereas Corollary 9.8 compares queues with different service times.

In fact, for deterministic service times we have the following simple relation. Consider two stationary $M/D/1$ queues with the FB discipline and a non-preemptive discipline $\pi \in \mathcal{D}$, e.g. FIFO. Let $Q_{\pi}$ and $Q_{FB}$ denote queue lengths. From the Pollaczek-Khinchin mean value formula, see e.g. Kleinrock [31], and Corollary 9.8, it follows that

$$EQ_{\pi} = \rho + \frac{\lambda^2 EB^2}{2(1-\rho)} = \rho + \frac{\rho^2}{2(1-\rho)} = \frac{2(1-\rho)p + \rho^2}{2(1-\rho)} = (1-\rho)\frac{(2-\rho)p}{2(1-\rho)^2} = (1-\rho)EQ_{FB}. \hspace{1cm} (9.12)$$

Relation (9.12) indicates the illness of the effect of using the FB discipline in the $M/D/1$ queue, instead of the FIFO discipline.
9.3 The stationary queue in heavy traffic

In this section the heavy traffic behaviour of the stationary $M/G/1$ FB queue is examined. Cohen [14] writes "Heavy traffic theory describes the behaviour of queueing systems with traffic intensities close to one. The main purpose of the theory is to derive asymptotic expressions for $\rho \uparrow 1$ of those relations which describe the behaviour of queueing systems". In this section we focus on the heavy traffic behaviour of the expected queue length in the stationary $M/G/1$ FB queue. For several classes of service-time distributions we find positive constants $\gamma$ such that $EQ = O((1 - \rho)^{-\gamma})$ as $\rho \uparrow 1$. It turns out that the values of $\gamma$ for bounded service times and unbounded service times are different. We start with a simple lemma.

**Lemma 9.9** Let $Q$ be the stationary queue length in an $M/G/1$ FB queue with an IMRL service-time distribution, defined in Definition 2.1. Then $EQ = O(1/(1 - \rho))$ as $\rho \uparrow 1$.

**Proof** For the stationary queue length $Q_{PS}$ in the $M/G/1$ PS queue with the same arrival rate, Theorem 6.3 gives $P(Q_{PS} = k) = (1 - \rho)\rho^k$. From Theorem 2.2 it follows that

\[(1 - \rho)EQ \leq (1 - \rho)EQ_{PS} = \sum_{k=0}^{\infty} (1 - \rho)k(1 - \rho)\rho^k \]
\[= \rho(1 - \rho)^2 \sum_{k=1}^{\infty} \frac{d\rho^k}{d\rho} = \rho(1 - \rho)^2 \frac{d}{d\rho} \frac{\rho}{1 - \rho} = \rho. \tag{9.13}\]

Hence $(1 - \rho)EQ = O(1)$ as $\rho \uparrow 1$. \qed

Consider a stationary $M/G/1$ FB queue with service-time distribution $F$ and arrival rate $\lambda$. Let $x_F$ be the end-point $F$, i.e. $x_F = \sup\{x : F(x) < 1\}$. It turns out that $Q$ shows different heavy traffic behaviour for $x_F = \infty$ and $x_F < \infty$. We first treat the case $x_F = \infty$. The following lemma and representation are needed.

**Lemma 9.10** Assume $EB < \infty$. Then

\[\rho - \rho(x) = \lambda \int_x^{\infty} (1 - F(t))dt.\]

**Proof** Integration by parts yields

\[\rho - \rho(x) = \lambda EB - \lambda E(B \land x) = \lambda \int_x^{\infty} (t - x)dF(t)\]
\[= -\lambda \int_x^{\infty} (t - x)d(1 - F(t)) = \lambda \int_x^{\infty} (1 - F(t))dt,\]
since, by Lemma 7.16, \((t - x)(1 - F(t)) \to 0\) for all \(x \in \mathbb{R}\) as \(t \to \infty\).

In proving the propositions in this chapter we make use of the following representation for \(EQ\), which is mentioned in Yashkov [69]. By Lemma 7.16, \(x(1 - F(x)) = o(1)\) for \(x \to \infty\). Since \(E(B \wedge x)^2 \leq xE(B \wedge x) \leq xEB\),

\[
(1 - F(x))E(B \wedge x)^2 \leq x(1 - F(x))EB = o(1). \quad x \to \infty. \quad (9.14)
\]

Using (9.14), we obtain from (2.9) by integration by parts

\[
EQ = -\lambda \int_0^\infty \left( \frac{\lambda E(B \wedge x)^2}{2(1 - \rho(x))^2} + \frac{x}{1 - \rho(x)} \right) d(1 - F(x))
\]

\[
= \int_0^\infty \left( \frac{\lambda(1 - F(x))}{1 - \rho(x)} + \frac{2\lambda^2 x(1 - F(x))^2}{(1 - \rho(x))^2} + \frac{\lambda^3(1 - F(x))^2 E(B \wedge x)^2}{(1 - \rho(x))^3} \right) dx
\]

(9.15)

Since \(1 - \rho < 1 - \rho(x)\) for every \(x\), from (9.15) it follows that

\[
(1 - \rho)EQ \leq \rho + 2\lambda^2 \int_0^\infty \frac{x(1 - F(x))^2}{\rho - \rho(x)} dx + \lambda^3 \int_0^\infty \frac{(1 - F(x))^2 E(B \wedge x)^2}{(\rho - \rho(x))^2} dx. \quad (9.16)
\]

**Lemma 9.11** Let \(x_F = \infty\). If \(EB^\alpha < \infty\) for some \(\alpha > 1\) and

\[
x(1 - F(x)) \sim c \int_{x}^\infty (1 - F(u)) du. \quad x \to \infty. \quad (9.17)
\]

for some \(c > 0\), then \(EQ = O(1/(1 - \rho))\) as \(\rho \uparrow 1\).

**Proof** Lemma 9.10, inequality (9.16) and (9.17) yield that for some constants \(c_1, c_2, c_3 \in \mathbb{R}\),

\[
EQ \leq c_1 + c_2 \int_0^\infty (1 - F(x))dx + c_3 \int_0^\infty x^{-2}E(B \wedge x)^2 dx.
\]

Since \(EB^\alpha < \infty\) for some \(\alpha > 1\), Lemma 7.16 yields that \(E(B \wedge x)^2 = O(x^{(2 - \alpha)^\wedge 0})\) for \(x \to \infty\). Hence \(\lim_{\rho \uparrow 1}(1 - \rho)EQ < \infty\), which finishes the proof. \(\Box\)

Note that for regularly varying tails of index \(\eta > 2\), the relation (9.17) follows by Karamata’s theorem, see for instance Bingham et al. [8].

**Lemma 9.12** Let \(x_F = \infty\). If \(1 - F(x) = o(\exp(-cx^\beta))\) for some \(\beta > 0\) as \(x \to \infty\), then \((1 - \rho)EQ = O(1/(1 - \rho))\) as \(\rho \uparrow 1\).
9.3 The stationary queue in heavy traffic

Proof From Lemma 9.10, inequality (9.16) and the assumption it follows that for some constants \(c_1, c_2, c_3 \in \mathbb{R}\),

\[
(1 - \rho)EQ \leq c_1 + c_2 \int_0^\infty x(1 - F(x))dx + c_3 \int_0^\infty (1 - F(x))^2 E(B \wedge x)^2 dx.
\]

Since \(\lim_{x \to \infty} E(B \wedge x)^2 < \infty\), there is a constant \(c_4 \in \mathbb{R}\) such that \((1 - \rho)EQ \leq c_4\) for all \(\rho\). This finishes the proof.

As we shall see next, \(EQ\) behaves differently in the case of a finite end-point \(x_F\).

Lemma 9.13 Let \(x_F < \infty\). If \(1 - F(x) \sim \alpha(x_F - x)^\beta\) for some \(\alpha, \beta > 0\) as \(x \to x_F\), then

\[
EQ = O((1 - \rho)^{-1 - 1/(\beta + 1)}).
\]

Proof First note that the assumption implies that

\[
\int_{x}^{x_F} (1 - F(t)) dt \sim \gamma(x_F - x)^{\beta + 1},
\]

where \(\gamma = \alpha/(\beta + 1)\). By Lemma 9.10, we then have for \(m, n \in \mathbb{N}\),

\[
\frac{(1 - F(x))^m}{(1 - \rho(x))^n} = \frac{(1 - F(x))^m}{(1 - \rho + \lambda \int_x^{x_F} (1 - F(t)) dt)^n} \sim \frac{\alpha(x_F - x)^{m\beta}}{(1 - \rho + \lambda \gamma(x_F - x)^{\beta + 1})^n}
\]

as \(x \to x_F\). Using the substitution \(x_F - x = y(1 - \rho)^{1/(\beta + 1)}(\lambda \gamma)^{-1/(\beta + 1)}\), we find that for some \(c, c_1, c_2, c_3 \in \mathbb{R}\),

\[
\int_0^{x_F} \frac{(1 - F(x))^m}{(1 - \rho(x))^n} dx = c_1 \int_0^{x_F} \frac{(x_F - x)^{m\beta}}{(1 - \rho + \lambda \gamma(x_F - x)^{\beta + 1})^n} dx
\]

\[
= c_2 \int_0^{x_F/(1 - \rho)^{1/(\beta + 1)}} \frac{(1 - \rho)^{m\beta/(\beta + 1)} y^{m\beta}}{(1 - \rho + (1 - \rho)^{y^{\beta + 1}})^n} dy (1 - \rho)^{1/(\beta + 1)}
\]

\[
= c_2 (1 - \rho)^{(m\beta + 1 - n\beta - n)/(\beta + 1)} \int_0^{x_F/(1 - \rho)^{1/(\beta + 1)}} \frac{y^{m\beta}}{(1 + y^{\beta + 1})^n} dy
\]

\[
= c_2 (1 - \rho)^{(m\beta + 1 - n\beta - n)/(\beta + 1)} \left( c_3 + \int_0^{x_F/(1 - \rho)^{1/(\beta + 1)}} y^{(m - n)\beta - n} dy \right) \quad (9.18)
\]

Using representation (9.15) and applying (9.18) with \((m, n) = (1, 1), (2, 2), (2, 3)\), there are constants \(c_4, c_5, c_6, c_7, c_8 \in \mathbb{R}\) such that

\[
EQ = c_5 \log \left( \frac{x_F}{(1 - \rho)^{1/(\beta + 1)}} \right) + c_6 (1 - \rho)^{-1/(\beta + 1)} \left( 1 - (1 - \rho)^{1/(\beta + 1)}/(cx_F) \right)
\]

\[
+ c_7 (1 - \rho)^{-(\beta + 2)/(\beta + 1)} \left( 1 - c_8 \left( \frac{(1 - \rho)^{1/(\beta + 1)}/x_F}{\beta + 2} \right) \right) + c_4. \quad (9.19)
\]
From (9.19) it follows that \( EQ = O((1 - \rho)^{-1-1/(\beta+1)}) \) as \( \rho \uparrow 1 \). This finishes the proof.

\[ \textbf{Example 9.14} \text{ Under the conditions of Lemma 9.13,} \]

\[ \lim_{\rho \uparrow 1} (1 - \rho)EQ = \infty. \]  

(9.20)

In particular, (9.20) holds if the service times are uniformly distributed on \((0,1)\).

We conclude that the expected queue length in the stationary M/G/1 FB queue in heavy traffic is smaller for infinite than for finite end-points of the service-time distribution. This is an example of a situation in which the FB discipline is more efficient for heavy-tailed service-time distributions, or light-tailed distributions with an infinite end-point, than for light-tailed distributions (with a finite end-point).

In the literature, a few heavy-traffic results exist for the FB queue. The survey paper Yashkov [69] quotes the following heavy-traffic limits from articles that appeared in Russian journals. For service-time distributions with tail \( 1 - F(x) \sim ax^b \exp(-cx) \), for some \( a > 0, b \geq 0, c > 0 \), the stationary queue length \( Q \) in the M/G/1 FB queue satisfies

\[ \lim_{\rho \uparrow 1} P(Q/EQ < x) = 1 - e^{-x}, \quad x \geq 0. \]

For the M/D/1 FB queue the Laplace transform of \( \lim_{\rho \uparrow 1} Q/EQ \) is given. Furthermore it is claimed that if the service times have a tail of the form \( ax^b, a > 0, b < -2 \), then the random variable \( Q/EQ \) has a non-exponential limiting distribution as \( \rho \uparrow 1 \).

\section{The cohort process}

In the FB queue customers cluster together in cohorts, groups of customers that have the same age. In this section we study the process describing this clustering. For that, we first describe \( Q(0, x) \), the number of customers younger than \( x \) in the stationary M/G/1 FB queue. Robert and Schassberger [54] proved the following theorem.

\textbf{Theorem 9.15 (Robert, Schassberger)} \textit{The process} \( Q(0, x) \) \textit{has independent increments}.

Hence \( Q(0, x) \) is a \textit{inhomogeneous} Poisson process with \textit{batch arrivals}, see Daley and Vere-Jones [19]. The mean increase \( \nu(x) = \frac{d}{dx}EQ(0, x) \) of the process \( Q(0, x) \) is found as follows.
Theorem 9.16 The mean increase \( \nu(x) \) of the process \( Q(0,x) \) is given by

\[
\nu(x) = \frac{\lambda(1 - F(x))}{1 - \rho(x)} + \frac{2\lambda^2 x(1 - F(x))^2}{(1 - \rho(x))^2} + \frac{\lambda^3 E(B \land x)^2(1 - F(x))^2}{(1 - \rho(x))^3}.
\]

Proof Let \( V_x \) be the sojourn time of a customer in the queue with generic service time \( B \land \tau \). Then

\[
EV_x = \int_0^x EV(u)dF(u) + EV(x)(1 - F(x)),
\]

(9.21)

where \( EV(u) \) is as in (2.11) and \( F \) is the distribution function of \( B \). Little’s law yields \( EQ(0,x) = \lambda EV_x \). Differentiation of the RHS of (9.21) then yields that \( \nu(x) = \lambda(1 - F(x))dEV(x)/dx \). The statement of the lemma follows from (2.15). □

The process \( Q(0,x) \) has independent increments, but it may have jumps of size larger than 1. A jump occurs at \( x \) when the cohort of age \( x \) contains more than one customer. In the following process, which is naturally linked to \( Q(0,x) \), all jumps have size 1.

By definition, for every \( x \) there is at most one cohort with age \( x \). Let \( K[a,b) \) be the number of cohorts in the stationary M/G/1 FB queue consisting of customers with age in \( [a,b) \). The cohort process may be derived immediately from the queue length process \( Q(0,x) \): each jump in the queue length process corresponds to a jump of size 1 in the cohort process \( K \). Combining this observation with Theorem 9.15, leads us to the following corollary.

Corollary 9.17 The cohort process \( K[0,x) \) is a (inhomogeneous) Poisson process.

Since \( K[0,x) \) has independent increments,

\[
P(K[a,b) = 0) = \frac{P(K[a,b) = 0, K[0,a) = 0)}{P(K[0,a) = 0)} = \frac{P(K[0,b) = 0)}{P(K[0,a) = 0)} = \frac{1 - \rho(b)}{1 - \rho(a)}.
\]

Here \( \rho(x) = \lambda E(B \land x) \). For \( 0 \leq a \leq b \) the expected number of points \( \mu([a,b)) \) in the interval \( [a,b) \) is given by

\[
\mu([a,b)) = -\log P(K[a,b) = 0) = \log(1 - \rho(a)) - \log(1 - \rho(b)).
\]

(9.22)

The intensity \( \mu(x) \) of the cohort process, given by \( \mu(x) = d\mu([0,x))/dx \), satisfies

\[
\mu(x) = \frac{d}{dx}(-\log(1 - \rho(x))) = \frac{\lambda(1 - F(x))}{1 - \rho(x)}.
\]

(9.23)

The expected size of a cohort with age \( x \), given that the cohort is non-empty, is equal to the quotient \( \nu(x)/\mu(x) \), where \( \nu(x) \) is the mean increase of \( Q(0,x) \) given
by Theorem 9.16. From (9.23) and Theorem 9.16 it follows that
\[
\frac{\nu(x)}{\mu(x)} = \frac{1 - \rho(x)}{\lambda(1 - F(x))} \lambda(1 - F(x)) \frac{dEV(x)}{dx} = (1 - \rho(x)) \frac{dEV(x)}{dx}.
\] (9.24)

In the remaining part of this section we derive the expression (9.24) in another way, giving insight in the queueing process. Consider the following situation. A 'tagged' customer with service time \(x\) or larger enters the stationary queue. Let \(H(x)\) be the number of customers that attain age \(x\) at the same time as the tagged customer. Hence \(H(x)\) is the size of the first cohort that reaches age \(x\) after the arrival of our tagged customer. In the same fashion as Lemma 2.13 we prove the next proposition about the mean of \(H(x)\).

**Proposition 9.18** If the service-time distribution \(F\) has no atom in \(x\), then
\[
EH(x) = 1 + \frac{2\lambda x(1 - F(x))}{(1 - \rho(x))} + \frac{\lambda^2 E(B \land x)^2(1 - F(x))}{(1 - \rho(x))^2}.
\]

**Proof** Assume without loss of generality that the tagged customer has service time larger than \(x\). By analysing the difference \(EV(x + \delta) - EV(x)\) as \(\delta \downarrow 0\), we find the expression for \(EH(x)\). This difference is the average time it takes for a customer with service time \(x + \delta\) to leave the queue, counting from the first moment his age is \(x\). The length of this period is the time the server spends serving three kinds of customers, namely other customers in the cohort, new customers, and customers in an older cohort.

The moment the age of the tagged customer reaches \(x\), there are no customers younger than \(x\) in the queue. However, before our tagged customer can leave the queue at the age \(x + \delta\), all the customers in his cohort have to attain age \(x + \delta\) or leave the queue before reaching that age. There are precisely \(H(x)\) customers in the cohort with age \(x\). If \(x\) is not an atom of \(F\), then each of these \(H(x)\) customers needs \(\delta - o(\delta)\) units of service to reach age \(x + \delta\) or leave the queue.

Secondly, by the nature of the FB discipline, the server gives priority to customers younger than \(x\). By the forgetfulness property of the exponential interarrival times, in a period of length \(EV(x + \delta) - EV(x)\), on average \(\lambda(EV(x + \delta) - EV(x))\) new customers arrive. To each of them an amount \(E(B \land (x + \delta))\) of service has to be given before our tagged customer can leave the queue. Thirdly, there is the possibility that the cohort of our tagged customer collides with a cohort with age in \((x, x + \delta)\). The probability of this event is \(O(\delta)\) as \(\delta \to 0\), by (9.23). Since the expected queue length is finite, the expected size of this cohort is finite as well. Hence the delay of this
type is $O(\delta^2)$.
Summarising,

$$
EV(x + \delta) - EV(x) = \delta(1 - o(1))EH(x) + \
+ \lambda E(B \wedge (x + \delta))(EV(x + \delta) - EV(x)) + O(\delta^2),
$$

and hence

$$
\delta EH(x) = (EV(x + \delta) - EV(x))(1 - \rho(x + \delta)) + o(\delta).
$$

Since (9.25) holds for all $\delta$, dividing by $\delta$ and taking the limit yields

$$
EH(x) = \lim_{\delta \to 0} \frac{1}{\delta} (1 - \rho(x + \delta))(EV(x + \delta) - EV(x)) + o(1)
= (1 - \rho(x)) \frac{dEV(x)}{dx},
$$

since $\rho(x) = \lambda E(B \wedge x)$ is continuous in $x$. The proof is completed by using (2.15). □

Hence if $x$ is not an atom of $F$, then $EH(x)$ has the same value as $\nu(x)/\mu(x)$, the expected size of a cohort in the stationary queue, given that the cohort is non-empty. The following corollaries hold by using Proposition 9.18 and taking limits.

**Corollary 9.19** Fix $x$ and suppose $F$ has no atom in $x$. Then

$$(1 - \rho(x))^2 EH(x) \to \lambda^2 E(B \wedge x)^2(1 - F(x)), \quad \rho(x) \uparrow 1.$$

We conclude that in heavy traffic, i.e. when the load $\rho$ is close to one, cohort sizes of non-zero cohorts tend to have large expectations. In Chapters 4, 5 and 6 we dwelled on the effect of increasing the variability of the service times. Here we devote a few words to the impact of more variability on $EH(x)$. Define the critical value $\lambda^*(x)$ by

$$
\lambda^*(x) = \sup\{\lambda : EH(x) < \infty\}.
$$

The function $h(u) = u \wedge x$ is concave. If the variability of $F$ increases in the convex order, see (4.1), then $E(B \wedge x)$ decreases for all $x$. Corollary (9.19) yields that $EH(x) = \infty$ if and only if $\rho(x) = \lambda E(B \wedge x) \geq 1$. Hence $\lambda^*(x)$ increases if the variability of $F$ increases. This again shows that the FB queue may behave better when the variability of the service-time distribution increases. Furthermore, since $1/E(B \wedge x)$ decreases in $x$, the size of older cohorts may have infinite expectation, while younger cohorts are still relatively small.
Corollary 9.20 By Lemma 2.17, \( \lim_{x \to \infty} EH(x) = 1 \) and
\[
\lim_{x \to 0} EH(x) = 1 + 2\lambda x + O(x^2).
\]
Note that this corollary agrees with the remark after Corollary 2.14.

9.5 A moment relation between cohort sizes and the queue length

Consider an M/G/1 queue with generic service time \( B \wedge x \) for some \( x > 0 \) and let\( F \) be the distribution function of \( B \). At time 0— the system is empty and at time 0 the first customer arrives. Let \( K_x \) denote the total number of customers in the busy period that have service time \( x \). Then \( K_x \) is independent of the service discipline. For the FB queue \( K_x \) has the following interpretation. At the end of the busy period, a batch of customers leaves the system. Then \( K_x \) is the number of customers in this batch that have age \( x \). Note that \( K_x \) may be zero. In this section we show the following relation between the mean queue length \( EQ \) and the second moment of \( K_x \).

Proposition 9.21 Let \( Q \) be the stationary queue length in the M/G/1 FB queue, and let \( K_x \) be as above. Then
\[
EQ = \lambda \int_0^\infty EK_x^2 \, dx. \tag{9.26}
\]

Pechinkin [44] uses the generating function \( \kappa(x, z) \) of \( K_x \) in his proof of Theorem 2.12. Balkema and Verwijmeren [6] obtain \( \kappa(x, z) \) by the method of the collective marks. Since this derivation is rather elegant, we repeat it below. After that the equality (9.26) is shown.

Proposition 9.22 (Pechinkin) The generating function \( \kappa(x, z) = EzK_z \) satisfies
\[
\kappa(x, z) = \int_0^x e^{-\lambda y(1-\kappa(x,z))} \, dF(y) + z(1 - F(x))e^{-\lambda x(1-\kappa(x,z))}, \quad 0 \leq z < 1.
\]

Proof All customers with service time \( x \) send a complaint to the manager of the system (since their service was cut off at level \( x \)). Complaints get lost, independently, with probability \( z \). Then \( \kappa(x,z) \) is the probability that all complaints get lost. Let \( B_1 \wedge x \) be the service time of the first customer. Since \( K_x \) is independent of the service discipline, we may serve the customers according to the FIFO discipline.
During the service of the first customer $K$ new customers enter, where $K$ is Poisson distributed with parameter $\lambda(B_1 \wedge x)$. All these customers start their own sub-busy period, at the end of which with probability $\kappa(x, z)$ no complaint reaches the manager. Using the fact that $E[t^K \mid B_1 = y] = e^{-\lambda y(1-t)}$ for $y \geq 0$, we find

$$
\kappa(x, z) = \int_0^\infty E[\kappa(x, z)^K \mid B_1 = x] dF(x) + zE[\kappa(x, z)^K \mid B_1 \geq x]P(B_1 \geq x)
$$

$$
= \int_0^x e^{-\lambda y(1-\kappa(x,z))} dF(y) + ze^{-\lambda y(1-\kappa(x,z))}(1 - F(x)),
$$

(9.27)

which finishes the proof.

Proof of Proposition 9.21 As stated in Theorem 2.12, the generating function of the stationary queue length $Q$ satisfies

$$
EzQ = (1 - \rho) \exp \left( - \int_0^\infty z \frac{\partial}{\partial z} v(t, z) dt \right),
$$

where $v(t, z)$ is the given by (2.8), as before. We may calculate

$$
EQ = \frac{\partial}{\partial z} Ez^Q \bigg|_{z=1} = -\frac{\partial}{\partial z} \int_0^\infty z \frac{\partial}{\partial z} v(t, z) \bigg|_{z=1} dt
$$

$$
= -\int_0^\infty \frac{\partial^2}{\partial z^2} v(t, z) - \frac{\partial}{\partial z} v(t, z) \bigg|_{z=1} dt.
$$

(9.28)

Comparing (2.8) and (9.27) reveals that $\lambda(1-\kappa(x, z)) = v(x, z)$. Hence the integrand in (9.28) is equal to

$$
\lambda \frac{\partial \kappa(x, z)}{\partial z} + \lambda \frac{\partial^2 \kappa(x, z)}{\partial z^2} \bigg|_{z=1} = \lambda EK_x + \lambda EK_x(K_x - 1) = \lambda EK_x^2.
$$

Proposition 9.26 follows from integrating w.r.t. $x$. \qed