Heavy Traffic Analysis for Continuous Polling Models

Dirk P. Kroese
Department of Applied Mathematics
University of Twente
P.O. Box 217, 7500 AE Enschede
The Netherlands

May 24, 2001

Abstract We consider a continuous polling system in heavy traffic. Using the relationship between such systems and age-dependent branching processes, we show that the steady-state number of waiting customers in heavy traffic has approximately a gamma distribution. Moreover, given their total number, the configuration of these customers is approximately deterministic.

Keywords: Queueing theory, continuous polling system, heavy traffic, queue length, age-dependent branching process, exponential approximation, gamma approximation, random measure, generating function.

AMS Subject Classifications (1991): Primary 60K25, 90B22, 60G55; Secondary 60G57, 60H05

1 Introduction

Polling systems are mathematical models for non-standard physical queueing systems in which clients are served in a cyclic order. These systems arise frequently in todays communication networks and manufacturing systems. A standard reference is Takagi [15]. For additional references see e.g. Boxma and Takagi [2]. Much insight in the behavior of polling systems is obtained through the analysis of *continuous* polling models, which have a "continuum" of waiting rooms. Examples may be found in Coffman and Gilbert [3], Fuhrmann and Cooper [7] and Kroese and Schmidt [11], [12].

To date, little is known about the behavior of discrete or continuous polling systems in heavy traffic, this in contrast to the well-developed heavy-traffic theory for classical queueing systems (see e.g. Whitt [17]). In Coffman et al. [4] it is shown that even for very simple polling systems it may be impossible to obtain heavy-traffic results in the usual sense. Heavy-traffic properties of polling systems are also mentioned in van der Mei [16].

A well-known result for the GI/G/1-queue in heavy traffic is that the steady-state waiting time distribution is approximately exponential. Moreover, the corresponding parameter depends only on the first two moments of the interarrival and service time distribution (see e.g. Kingman [9]). A similar *exponential approximation* holds for the queuelength distribution (see e.g. Szczotka [14]). Motivated by this relatively simple behavior, we consider a continuous polling system in heavy traffic. The close relationship between polling systems and branching processes serves as a basis for our investigations. This relationship is recognized and used frequently in the literature, see for example Resing [13] and references there, but still seems not to have been fully exploited.

It turns out that in heavy traffic the steady-state number of customers in the system has approximately a gamma distribution, depending only on a few (known) parameters. Moreover, it is shown that the configuration of waiting customers is approximately deterministic, given their total number.

In the next section we briefly review the queueing system under consideration. In Section 3 we derive the "gamma-approximation" for this system. For simplicity, we restrict ourself to heavy-traffic approximations for stable queues. Generalizations are left to the reader. Finally, Section 4 discusses some (heavy-traffic) results for the expected steady-state number of customers.

2 Preliminaries

In this section we briefly review the continuous polling model of Coffman and Gilbert [3] and Kroese and Schmidt [11]. Throughout this paper $(\Omega, \mathcal{H}, \boldsymbol{P})$ denotes the probability space in the background, with corresponding expectation symbol $\boldsymbol{E}$. $\mathcal{B}[0,1]$ denotes the set of positive measurable functions on $[0,1]$. The class of continuously differentiable functions on [0,1] is denoted by $C^1[0,1]$. C_K^+ denotes the class of positive continuous functions on [0,1] whose support is compact. We will frequently write μf for the integral of a function f with respect to a (random) measure μ. Finally, $X_n \xrightarrow{\mathrm{d}} X$ denotes convergence in distribution of the sequence of random variables (X_n) to the random variable X. Basic definitions and results on random (counting) measures can be found for example in Daley and Vere-Jones [5]. We refer to Athreya and Ney [1] and Jagers [8] for details on branching processes.

Consider a queueing system in which customers arrive according to a homogeneous Poisson process with rate a on a ring with circumference 1. Incoming customers take their positions on the ring according to a uniform distrubution, and wait there to be served by a server who travels on the ring at constant speed α^{-1} (in one direction). The customers are served in the order in which they are encountered by the server. During a service the server does not travel. The service time distribution function is F, with first moment e_1 and finite second moment e_2. Starting from an empty system at time 0, let W_t denote the random measure on [0,1] representing the positions of waiting customers at time t *relative* to the position of the server. Similarly, let the random measure Q_t denote the configuration of waiting customers when the server has been *traveling* for t units of time. And finally, let Q_t^0 denote the configuration of waiting customers when the server has been *busy* for t units of time. It is shown in Kroese and Schmidt [11] that the measure-valued processes $(W_t), (Q_t)$ and (Q_t^0) are *regenerative* when the traffic intensity ae_1 is less than 1. And in this case there exist therefore limiting random measures (on [0,1]) W, Q and Q^0 to which $(W_t), (Q_t)$ and (Q_t^0) converge in distribution. We can interpret Q as the random measure of waiting customers relative to the server, in the stationary situation, at traveling epochs. Similarly Q^0 and W correspond to the stationary configuration of customers at service epochs and at "random" epochs, respectively.

The laws of W, Q and Q^0 are completely specified by the following three propositions (see Kroese and Schmidt [11] for proofs):

Proposition 1 The law of W is a mixture of that of Q and Q^0:

$$\boldsymbol{E}\mathrm{e}^{-Wf} = (1-ae_1)\boldsymbol{E}\mathrm{e}^{-Qf} + ae_1\boldsymbol{E}\mathrm{e}^{-Q^0 f}, \quad \text{for all } f \in \mathcal{B}[0,1].$$

Proposition 2 The Laplace functional of Q satisfies

$$\boldsymbol{E}\mathrm{e}^{-Qf} = \exp -a\alpha \int_0^\infty dt\,\{1-K_{a,f}(t)\}, \quad \text{for all } f \in \mathcal{B}[0,1],$$

where $K_{a,f}$ is the solution to

$$\begin{aligned} K(t) &= \mathbf{1}_{[0,1]}(t)\int_0^t dx\,\varphi(K(t-x)) + \mathbf{1}_{[0,1]}(t)\int_t^1 dx\,\mathrm{e}^{-f(x-t)} \\ &+ \mathbf{1}_{(1,\infty)}(t)\int_0^1 dx\,\varphi(K(t-x)). \end{aligned} \tag{1}$$

Here $\mathbf{1}_I$ denotes the indicator function of an interval I and φ is defined by

$$\varphi(z) = \int_0^\infty \mathrm{e}^{-a(1-z)x}\, dF(x),\ z \in [0,1]. \tag{2}$$

Proposition 3 The law of Q^0 is completely determined by the law of Q, because

$$\boldsymbol{E}\mathrm{e}^{-Q^0 f} = \frac{\boldsymbol{E}\mathrm{e}^{-Qf}\,(\alpha^{-1}Qf' - \beta)}{ae_1(\beta+\gamma)/(1-ae_1)}, \quad \text{for all } f \in C^1[0,1], \tag{3}$$

where

$$\beta = a\int_0^1 dx\,(1-\mathrm{e}^{-f(x)}),$$

$$\gamma = (1-\mathrm{e}^{f(0)})\frac{\beta L_F(\beta)}{1-L_F(\beta)},$$

and L_F is the Laplace-Stieltjes transform of F.

In the proof of Proposition 2, the fact is used that the measure valued process (Q_t) is closely related to the following *particle system* on the strip $I\!R_+ \times [0,1]$ (time $\times$ position): A first generation particle starts at time $t=0$ from position U, where U is uniformly distributed on $[0,1]$. This particle moves towards 0 with unit speed, and when it hits 0, it dies, but at the same time J new particles are born, at positions independently and uniformly distributed on [0,1], where J has generating function φ given in (2). Notice that φ is exactly the generating function of the number of customers that arrive during a service period. Let L_t denote the random measure on [0,1] whose atoms are formed by the particles that are alive at time t. For each $0 \le a \le 1/e_1$ and $f \in \mathcal{B}[0,1]$, the function $K_{a,f}$ in Proposition 2 and the random measure L_t are related via

$$K_{a,f}(t) = \boldsymbol{E}\mathrm{e}^{-L_t f},\ \ t \ge 0. \tag{4}$$

Notice that the process $(|L_t|) := (L_t[0,1])$ is an age-dependent (or even a Bellman-Harris) branching process. Each individual has a uniformly distributed lifetime, and the offspring generating function is φ. And (1) is a generalized version of equation (6.3.3) of Jagers [8]. Similarly, it is not difficult to see that the process $(|Q_t|) := (Q_t[0,1])$ is the corresponding Bellman-Harris branching process with Poisson *immigration*.

Although Propositions 1–3 specify the distributions of W and Q and Q^0 completely, the *actual* distributions of these measures are not known in the general case. Even the distribution of $|Q|$ is not known. This is another motivation for considering heavy-traffic approximations. However, for the *constant* service time case the Laplace functional of Q is know explicitly. Specifically,

$$\boldsymbol{E}\mathrm{e}^{-Qf} = \mathrm{e}^{-c_f}\left(\frac{1-ae_1}{1-ae_1\int_0^1 dy\,\mathrm{e}^{-h(y)}}\right)^{\alpha/e_1}, \quad \text{for all } f \in \mathcal{B}[0,1], \tag{5}$$

where $c_f = \alpha a \int_0^1 dx\,(1-x)(1-\mathrm{e}^{-f(x)})$ and $h(y) = ae_1 \int_0^y dx\,(1-\mathrm{e}^{-f(x)})$.

We finally give some results on the moment measures of Q and Q^0. An easy consequence of (3) is that the mean measure of Q satisfies

$$\boldsymbol{E}Q(dx) = \frac{\alpha a}{1-ae_1}\,(1-x)\,dx,\;\; x \in [0,1]. \tag{6}$$

With more effort we can find second-moment results. For example, for every $f \in \mathcal{B}[0,1]$ we have

$$\begin{aligned} \text{var } Qf &= \frac{\alpha a}{1-ae_1}\int_0^1 dx\,(1-x)f^2(x) + \frac{\alpha a^3 e_2}{1-ae_1}\int_0^1 dy\left(\int_0^y dx\, f(x)\right)^2 \\ &+ \frac{\alpha a^4 e_1 e_2}{(1-ae_1)^2}\left(\int_0^1 dx\,(1-x)f(x)\right)^2. \end{aligned} \tag{7}$$

In the next section we need also the mean measure of Q^0. The following lemma is new.

Lemma 1 The mean measure of Q^0 is given by

$$\boldsymbol{E}Q^0(dx) = \delta_0(dx) + \frac{a\,dx}{2(1-ae_1)}\left\{2\alpha(1-x) + e_2(e_1^{-1} + a - 2ax)\right\},\;\; x \in [0,1],$$

where δ_0 is the Dirac measure at 0.

Proof This follows basically from (6), (7) and Proposition 3. Namely, a Taylor-expansion of (3) yields,

$$\begin{aligned} \boldsymbol{E}Q^0 f &= \Big(ae_1^{-1}(e_1^2 - \frac{e_2}{2})f(0)\int_0^1 dx\, f(x) + \frac{1-ae_1}{\alpha a}\boldsymbol{E}QfQf' \\ &- \alpha a \int_0^1 dx\, f(x)\int_0^1 dx\,(1-x)f(x) - \frac{1}{2}f^2(0) \\ &- \frac{1}{2}\int_0^1 dx\, f^2(x)\Big) \Big/ \Big(e_1 a \int_0^1 dx\, f(x) - f(0)\Big), \end{aligned}$$

where $\boldsymbol{E}QfQf'$ can be calculated from (6) and (7), because

$$\begin{aligned} \boldsymbol{E}QfQf' &= \frac{1}{2}\Big(\text{ var } Q(f+f') - \text{ var } Qf - \text{ var } Qf' \\ &+ \{\boldsymbol{E}Q(f+f')\}^2 - \{\boldsymbol{E}Qf\}^2 - \{\boldsymbol{E}Qf'\}^2\Big). \end{aligned}$$

The easiest way to proceed is to consider for each $x \in (0,1]$ the continuous "trapezoid" function f which is 1 in the interval $[x-dx, x+dx]$, 0 on $[0, x-dx-\epsilon] \cup [0, x+dx+\epsilon]$ and linearly in- or decreasing elsewhere. For such a function, $\boldsymbol{E}QfQf'$ and hence $\boldsymbol{E}Q^0 f$ can be explicitly calculated. Now let $\epsilon \to 0$ and $du \to 0$, and conclude that on (0,1] the measure $\boldsymbol{E}Q$ has a density with respect to the Lebesgue-measure. Similarly, by considering the continuous piecewise linear function that is 1 at 0 and 0 in $[\epsilon, 1]$, we find that $\boldsymbol{E}Q$ has an atom at 0 of size 1. $\square$

Remark 1 Since we will not make further use of the random variables W_t, Q_t and Q^0_t, we will *redefine* the symbols W_n, Q_n and Q^0_n in the next section. Moreover, in order to simplify the notation, we will take from now on $e_1 = 1$, and write b for e_2.

3 Gamma approximation

In this section we analyze the behavior of the continuous polling in heavy traffic. To this end, consider a sequence of traffic intensities (a_n) increasing to 1 (remember that the expected service time is 1). To each pair (a_n, F) there correspond unique "configuration measures" W, Q and Q^0 on [0,1], defined in the previous section, which we will denoted by W_n, Q_n and Q^0_n, respectively (see also Remark 1).

We start with bounds on the function $K_{a,f}$, defined through (1). Notice that for each pair (a, F) the distributions of the configuration measures depend ultimately on this function.

Lemma 2 For every $f \in \mathcal{B}[0,1]$, with $||f|| := \sup_x f(x) < \infty$ and $f \geq 1$, and for each $a < 1$, we have $K_{a,f}(t) \uparrow 1$ as $t \uparrow \infty$ and

$$\int_0^1 dx\, \mathrm{e}^{-f(x)} \leq K_{a,f}(t) \leq 1, \quad \text{for all } t \geq 0.$$

Proof Take $a < 1$ and let f be such that $f(x) \geq 1$ for all x. By Jensens' Inequality,

$$\varphi(z) \geq \mathrm{e}^{-a(1-z)} \geq e^{-1}, \quad \text{for all } 0 \leq z \leq 1.$$

Writing (1) in differential form:

$$K'(t) = \begin{cases} \varphi(K(t)) - \mathrm{e}^{-f(1-t)}, & 0 \leq t < 1, \\ \varphi(K(t)) - \varphi(K(t-1)), & t > 1, \end{cases}$$

we see that the derivative of $K_{a,f}$ is positive in [0,1]. And because φ is increasing (it is a generating function), $K'_{a,f}$ is also positive in $(1, \infty)$. Since f is positive, we have by (4)

$$\boldsymbol{E}\mathrm{e}^{-|L_t|\cdot||f||} \leq \boldsymbol{E}\mathrm{e}^{-L_t f} = K_{a,f}(t) \leq 1, \quad \text{for all } t \geq 0.$$

Since $(|L_t|)$ is an age-dependent sub-critical branching process with maximum lifespan 1, $|L_t| \to 0$ with probability 1. This shows that $\lim_{t\to\infty} K_{a,f}(t) = 1$. Finally, the lower bound for $K_{a,f}$ follows from the fact that $K_{a,f}(0) = \boldsymbol{E}\mathrm{e}^{-L_0 f} = \int_0^1 dx\, \mathrm{e}^{-f(x)}$. □

The next theorem shows that, in heavy traffic, $(1-a_n)Q_n$ is approximately of the form $|Q^*|\, 2(1-x)\, dx$, where $|Q^*|$ has a Gamma$(\alpha/b, 2/b)$-distribution. We can interpret this as follows: the total number of waiting customers in the stationary situation, given that the server is busy, has in heavy traffic approximately a gamma distribution. And, moreover, given this total number, the "density" of waiting customers relative to the server is (approximately) deterministic and linearly decreasing.

Theorem 1 The sequence of random measures $((1-a_n)Q_n)$ converges in distribution to a random measure Q^* on [0,1] with Laplace functional

$$\boldsymbol{E}\mathrm{e}^{-Q^* f} = \left(\frac{1}{1 + b\int_0^1 dx\,(1-x)f(x)} \right)^{\alpha/b}, \quad \text{for all } f \in \mathcal{B}[0,1]. \tag{8}$$

Proof In order to prove the theorem, it suffices to show (see Remark 3) that

$$\lim_{n\to\infty} \boldsymbol{E}\mathrm{e}^{-(1-a_n)Q_n f} = \boldsymbol{E}\mathrm{e}^{-Q^* f}, \tag{9}$$

for all all $f \in C_K^+$ that are greater equal to 1. For the rest of the proof, we will only consider such functions f.

Let K_n be the solution of (1) with a_n substituted for a and $(1-a_n)f$ substituted for f; in other words $K_n := K_{a_n,(1-a_n)f}$. By Proposition 2

$$\begin{aligned} \boldsymbol{E}\mathrm{e}^{-(1-a_n)Q_n f} &= \exp - a_n\alpha \int_0^\infty dt\,\{1 - K_n(t)\} \\ &= \exp - a_n\alpha \int_0^\infty dt\, g_n(t), \end{aligned}$$

where g_n is defined by

$$g_n(t) := \frac{1 - K_n\left(t/(1-a_n)\right)}{1 - a_n}, \quad t \geq 0.$$

Since f satisfies the conditions of Lemma 2, we infer from this lemma that for each n, g_n is a bounded decreasing function with

$$0 \leq g_n(t) \leq \int_0^1 dx\, f(x), \quad t \geq 0.$$

If we can show that (g_n) converges pointwise on $(0,\infty)$ to a function g with

$$\int_0^\infty dt\, g(t) = b^{-1}\log(1 + b\int_0^1 dx\,(1-x)f(x)), \tag{10}$$

then the theorem is proved, by the bounded convergence theorem.

We proceed in the same spirit as Jagers [8, Section 7.4]. Let φ_n denote the generating function in (2) with $a := a_n$. A Taylor-expansion of φ_n around 1 yields,

$$\varphi_n(z) = 1 - a_n(1-z) + \frac{a_n^2 b\,(1-z)^2}{2} - \varepsilon_n(z)(1-z)^2. \tag{11}$$

Let us denote the uniform probability measure on $[0, 1-a_n]$ by L_n. It follows from (1) and (11) that

$$\begin{aligned} g_n(t) &= \int_0^t L_n(dy)\, \{a_n g_n(t-y) - \frac{a_n^2 b}{2}\,(1-a_n) g_n^2(t-y)\} && (12)\\ &+ \mathbf{1}_{[0,1]}\left(\frac{t}{1-a_n}\right) \int_t^1 L_n(dy)\, f\left(\frac{y-t}{1-a_n}\right) && (13)\\ &+ (1-a_n) \int_0^t L_n(dy)\, \varepsilon_n(K_n\left(\frac{t-y}{1-a_n}\right))\, g_n^2(t-y) && (14)\\ &- \mathbf{1}_{[0,1]}\left(\frac{t}{1-a_n}\right) \int_t^1 L_n(dy)\, \tilde{\varepsilon}_n\left(\frac{y-t}{1-a_n}\right), && (15)\end{aligned}$$

where $\tilde{\varepsilon}_n$ is defined by

$$\mathrm{e}^{-f(x)(1-a_n)} = 1 - f(x)(1-a_n) + \tilde{\varepsilon}_n(x)(1-a_n).$$

We now take the Laplace-Stieltjes transforms (LSTs) of all terms in (12)–(15). The LST of a function h is denoted by $\hat{h}$. We obtain for every $s \geq 0$,

$$\begin{aligned} \widehat{g_n}(s) &= \frac{1-\mathrm{e}^{-s(1-a_n)}}{s(1-a_n)}\, \{a_n\, \widehat{g_n}(s) - \frac{a_n^2 b}{2}\,(1-a_n)\, \widehat{g_n^2}(s)\} \\ &+ \int_0^1 dx\, f(x)\,(1-\mathrm{e}^{-s(1-x)(1-a_n)}) \;+\; (1-a_n) r_n(s), \end{aligned}$$

where $(1-a_n)r_n$ is the LST of the sum of the terms in (14) and (15). By Lemma 2, K_n is uniformly bounded from below by $1-(1-a_n)\int_0^1 dx\, f(x)$. Moreover, it will be shown in Lemma 3 that $\varepsilon_n(\delta_n) \to 0$ as $n \to \infty$, for any positive sequence (δ_n), increasing to 1. Also, $\tilde{\varepsilon}_n(x) \to 0$, uniformly for $x \in [0,1]$. Consequently $r_n(s)$ converges to 0 uniformly in s, as $n \to \infty$. It follows that we can rewrite the equation above as

$$\widehat{g_n}(s)\, \left(1 + \frac{a_n s}{2}\right) = -\frac{a_n^2 b}{2}\, \widehat{g_n^2}(s) + s \int_0^1 dx\,(1-x) f(x) + \eta_n(s), \qquad (16)$$

for every $s \geq 0$, where $\eta_n(s) \to 0$ uniformly in s, as $n \to \infty$.

Next, by Lemma 2, (g_n) is a sequence of bounded decreasing functions. We infer from (a straightforward modification of) Helly's Selection Theorem that every subsequence of (g_n) has a further subsequence that converges to some limit g weakly. If we can show that there is only one limit g possible, we have proved that (g_n) converges weakly to g. Suppose therefore that (n_k) is some subsequence along which (g_n) converges to some limit g. By (16) and the continuity of LSTs, g satisfies

$$\frac{b}{2}\, \widehat{g^2}(s) + \widehat{g}(s)\, \left(1 + \frac{s}{2}\right) - s\, c = 0, \qquad (17)$$

with $c = \int_0^1 dx\,(1-x)f(x)$. Lemma's (7.4.3.) and (7.4.4) of Jagers [8] show that for a given positive "initial value" $g(0+)$, the integral equation above has exactly one *bounded* solution. The value of $g(0+)$ follows from (17) by a Tauberian Theorem (Feller [6, p. 442]), which says that the initial behavior of g is completely determined by the asymptotic behavior of $\widehat{g}(s)$ as $s \to \infty$. Specifically, the fact that here $\lim_{s\to\infty} \widehat{g}(s) = 2c$, implies that also $g(0+) = 2c$. Hence, by Lemma (7.4.3) of Jagers [8],

$$g(t) = \frac{2c\,\mathrm{e}^{-2t}}{1 - bc(e^{-2t} - 1)}, \quad t > 0. \tag{18}$$

Since g satisfies exactly the requirement of (10), this concludes the proof. □

Remark 2 The fact that the heavy-traffic limiting distribution is not exponential should not come as a surprise. It is well-known that the limiting distribution of the waiting time in the M/G/1-queue can have a variety of distributions, depending on the *service discipline* of the queue, see e.g. Kingman [10]. The exponential approximation tends only to appear for FIFO disciplines. Indeed, in polling systems the queue discipline is quite complicated and is usually neither FIFO, LIFO or ROS (random order of service). Furthermore, gamma distributions arise often as limiting distributions of critical branching processes. Therefore, in this respect, the appearance of the gamma approximation for polling systems is (in hindsight) not unexpected. Notice also that in the constant service time case (8) easily follows from (5).

The following lemma was used in the proof of Theorem 1.

Lemma 3 Let ε_n be remainder term in (11) and let (δ_n) be a positive sequence that increases to 1. Then

$$\lim_{n\to\infty} \varepsilon_n(\delta_n) = 0. \tag{19}$$

Proof It is easy to check (see Athreya and Ney [1, p. 63]) that ε_n is a positive decreasing function, with $\varepsilon_n(z) \downarrow 0$ as $z \uparrow 1$. Moreover, from another look at Taylor-expansion of φ_n around 1, we obtain that for every $z \in [0,1)$,

$$\frac{1}{2}\,a_n^2 b - \varepsilon_n(z) = \frac{1}{2}\,\varphi_n''(\xi),$$

for some $\xi \in [z,1)$. Since φ_n'' is increasing and $\varphi_n''(1) = a_n^2 b$, the last equation implies

$$2\,\varepsilon_n(z) \le \varphi_n''(1) - \varphi_n''(z) = \le \boldsymbol{E} J_n^2(1 - z^{J_n}),$$

where J_n is a random variable with generating function φ_n. Let J be a random variable with a generating function given by (2) with $a = 1$. Since $J_n \xrightarrow{\mathrm{d}} J$ and $\boldsymbol{E}J_n^2 \to \boldsymbol{E}J^2 = b$, the sequence $(J_n^2(1 - \delta_n^{J_n}))$ converges in expectation to 0, and (19) follows. □

Remark 3 The random measures that we encounter here can be seen as random variables with values in the measurable space $(E, \mathcal{E})$, where E is the space of all totally finite measures on $\mathcal{B}[0,1]$ endowed with the Prohorov distance d and Borel σ-algebra $\mathcal{E}$. The topology generated by d coincides with the topology of weak convergence:

$$d(\mu, \mu_n) \to 0 \iff \mu_n f \to \mu f,$$

for all bounded continuous functions on [0,1]. Moreover, E with this topology is a complete separable metric space (c.s.m.s.).

Suppose $X_1, X_2, \ldots$ and X are (totally finite) random measures on [0,1]. Let P_n be the distribution of X_n, $n = 1, 2, \ldots$ and let P be the distribution of X. Let ω_n and ω be the Laplace functionals of X_n and X (or P_n and P), respectively: $\omega_n(f) = \boldsymbol{E}\mathrm{e}^{-X_n f}$ and $\omega(f) = \boldsymbol{E}\mathrm{e}^{-Xf}$, for $f \in \mathcal{B}[0,1]$. The *continuity theorem* for random measures states that $X_n \xrightarrow{\mathrm{d}} X$, or equivalently, that (P_n) converges weakly to P, if and only if

$$\omega_n(f) \to \omega(f), \quad \text{for all } f \in C_K^+. \tag{20}$$

See e.g. Daley and Vere-Jones [5, Proposition 9.1.VII.].

In Theorem 1 the fact is used that (20) holds if and only if it holds for functions in C_K^+ that are greater or equal to 1. This follows from the *translation principle*, cf. Feller [6, p. 433]), applied to random measures: Assume that $\omega_n(f) \to \omega(f)$, for $f \geq 1$. Let $\mathbf{1}$ be the function $x \mapsto 1$. The functional $h \mapsto \omega_n(h + \mathbf{1})/\omega_n(\mathbf{1})$, $h \in \mathcal{B}[0,1]$ is the Laplace functional of the distribution $P_n^{\#}(d\mu) := P_n(d\mu)\,\mathrm{e}^{-|\mu|}/\omega_n(\mathbf{1})$ on $(E, \mathcal{E})$, where $|\mu| = \mu[0,1]$. By the continuity theorem there exists a unique distribution $P^{\#}$ on $(E, \mathcal{E})$, such that $(P_n^{\#})$ converges weakly to $P^{\#}$. This implies that (P_n) converges weakly to the distribution $\omega(\mathbf{1})\,\mathrm{e}^{|\mu|}P^{\#}(d\mu)$, which has Laplace functional ω and is therefore equal to P. □

Next, we analyze the distributions of the steady-state configuration of waiting customers at service epochs and at "random " epochs. Specifically, we study the behavior of the distributions of the random measures Q^0 and W in heavy traffic. We consider again the simplest situation where the service time distribution F is fixed and where we have a sequence of traffic intensities $(a_n) \uparrow 1$. Let Q_n^0 and W_n denote the corresponding measures Q^0 and W. Also define $Q_n^* := Q_n(1 - a_n)$, and let Q^* be the limiting distribution of (Q_n^*), as in Theorem 1.

Theorem 2 The sequences $(W_n(1 - a_n))$ and $(Q_n^0(1 - a_n))$ converges in distribution to the same random measure W^* with Laplace functional given by

$$\boldsymbol{E}\mathrm{e}^{-W^* f} = \left(\frac{1}{1 + b\int_0^1 dx(1-x)\, f(x)}\right)^{\frac{\alpha}{b}+1}, \quad f \in \mathcal{B}[0,1]. \tag{21}$$

Proof We first consider convergence of the sequence $(Q_n^0(1-a_n))$. For clarity write X_n for $Q_n^0(1-a_n)$ and X for W^*. In order to show that (X_n) converges in distribution to X it suffices to show that

$$\lim_{n\to\infty} \boldsymbol{E}\mathrm{e}^{-X_n f} = \boldsymbol{E}\mathrm{e}^{-Xf}, \quad \text{for all } f \in C_K^+. \tag{22}$$

We first show that it suffices to check (22) only for functions in $C^1[0,1]$. To see this, let f be a function in C_K^+ and let (f_k) be a sequence of uniformly bounded functions in $C^1[0,1]$ such that $\int_0^1 dx\,|f_k(x) - f(x)| \to 0$, as $k \to \infty$. Such a sequence is always to find. Assume that (22) holds for each f_k. Now,

$$\begin{aligned} &|\boldsymbol{E}\mathrm{e}^{-X_n f} - \boldsymbol{E}\mathrm{e}^{-Xf}| \\ &\leq\ |\boldsymbol{E}\mathrm{e}^{-X_n f} - \boldsymbol{E}\mathrm{e}^{-X_n f_k}| + |\boldsymbol{E}\mathrm{e}^{-X_n f_k} - \boldsymbol{E}\mathrm{e}^{-X f_k}| + |\boldsymbol{E}\mathrm{e}^{-X f_k} - \boldsymbol{E}\mathrm{e}^{-Xf}| \\ &\leq\ \mu_n|f_k - f| + |\boldsymbol{E}\mathrm{e}^{-X_n f_k} - \boldsymbol{E}\mathrm{e}^{-X f_k}| + \mu|f_k - f|, \end{aligned} \tag{23}$$

where μ_k is the mean measure of X_n and μ the mean measure of X. By assumption, the second term in (23) goes to 0 as $n \to \infty$ (k fixed). The measure μ_n is (by Lemma 1) given by

$$\mu_n(dx) = (1-a_n)\delta_0(dx) + \frac{a_n\,dx}{2}\,\{2\alpha(1-x) + b(1 + a_n - 2a_n x)\}\,,\ \ x \in [0,1],$$

and from (21) it follows that

$$\mu(dx) = (\alpha + b)(1-x)\,dx,\ \ x \in [0,1].$$

Since the f_k are uniformly bounded and converge to f in L^1-sense, the limit (for $n \to \infty$) of the first and third term in (23) can be made arbitrarily small by choosing k large enough. This shows that (22) is true for all $f \in C_K^+$ if it is true for all $f \in C^1[0,1]$.

Secondly, by Proposition 3, we have for all $f \in C^1[0,1]$,

$$\boldsymbol{E}\mathrm{e}^{-X_n f} = \frac{\boldsymbol{E}\mathrm{e}^{-Q_n^* f}\,(\alpha^{-1}Q_n^* f' - \beta_n)}{a_n(\beta_n + \gamma_n)/(1-a_n)},$$

where

$$\beta_n = a_n \int_0^1 dx\,(1 - \mathrm{e}^{-f(x)(1-a_n)}),$$

$$\gamma_n = (1 - \mathrm{e}^{f(0)(1-a_n)})\frac{\beta_n L_F(\beta_n)}{1 - L_F(\beta_n)},$$

and L_F is the Laplace-Stieltjes transform of F. Since we can view Q_n^* and Q^* as random variables taking values in the c.s.m.s. (E, d) of Remark 3, Skorohod's Representation Theorem is in force, which elevates convergence in distribution to almost sure convergence. Specifically, because (Q_n^*) converges in distribution to Q^*, there exists a probability space $(\Omega', \mathcal{H}', \boldsymbol{P}')$ and random measures Y_n and

Y on [0,1], with the same distribution as Q_n^* and Q^*, respectively, such that (Y_n) converges $\boldsymbol{P}'$-almost surely to Y. Consequently, since f' is bounded, $(\mathrm{e}^{-Y_n f} Y_n f')$ converges $\boldsymbol{P}'$-almost surely to $\mathrm{e}^{-Yf} Y f'$. Moreover, the sequence $(\mathrm{e}^{-Y_n f} Y_n f')$ is uniformly integrable, because $\boldsymbol{E} Q_n^* f'$ converges to $\alpha \int_0^1 dx\,(1-x) f'(x)$. Hence, as $n \to \infty$,

$$
\begin{aligned}
\boldsymbol{E}\mathrm{e}^{-Q_n^* f} Q_n^* f' &= \boldsymbol{E}'\mathrm{e}^{-Y_n f} Y_n f' \to \boldsymbol{E}'\mathrm{e}^{-Yf} Y f' = \boldsymbol{E}\mathrm{e}^{-Q^* f} Q^* f' \\
&= \alpha \left(\frac{1}{1+cb}\right)^{\frac{\alpha}{b}+1} \left(\int_0^1 dx\, f(x) - f(0)\right),
\end{aligned}
$$

where $c = \int_0^1 dx(1-x)\, f(x)$. The last equation above follows from the fact that Q^* has the same distribution as the random measure $Z\, 2(1-x)\, dx$ on [0,1], where the random variable Z has a Gamma$(\alpha/b, 2/b)$-distribution. Also,

$$
\lim_{n\to\infty} (\beta_n + \gamma_n)/(1-a_n) = \int_0^1 dx\, f(x) - f(0).
$$

This shows that the Laplace functional of (X_n) converges to that of $X = W^*$ for all continuously differentiable functions, which had to be shown. The fact that $(W_n(1-a_n))$ converges also in distribution to W^* now follows simply from Proposition 1. □

Remark 4 Theorem 2 shows that when we observe the system at service epochs or at random epochs, the limiting distribution of the number of waiting customers is again approximately a gamma distribution. Notice that when the cycle time α of the server is 0, we arrive at the exponential approximation for the M/G/1-queue.

4 Expectation

We conclude with a somewhat heuristic study of the expectations of the (total) number of waiting customers.

Let $\{a_n\}$ be a sequence of arrival intensities increasing to 1, and let Q_n^* and Q^* be as in the previous section. From (6) it is obvious that for all $f \in \mathcal{B}[0,1]$,

$$
\lim_{n\to\infty} \boldsymbol{E} Q_n^* f = \boldsymbol{E} Q^* f.
$$

It follows that the mean measure of Q_n^* converges weakly to that of Q^*. Similarly, by (7)

$$
\lim_{n\to\infty} \boldsymbol{E}\,(Q_n^* f)^2 = \boldsymbol{E}\,(Q^* f)^2 .
$$

In order to investigate the convergence of higher moments, we consider the third moment of $|Q| := Q([0,1])$, for some arbitrary arrival rate a. As is explained in Kroese and Schmidt [11], Proposition 2 leads directly to computable expressions for the moments of $|Q|$, provided that the corresponding moments of the service time distribution exist. In particular, for the third moment, we have

$$\begin{aligned} \boldsymbol{E}|Q|^3 &= \left(\alpha a \int_0^\infty ds\, k(s)\right)^3 \\ &+ 3\left(\alpha a \int_0^\infty ds\, k(s)\right)\left(\alpha a \int_0^\infty ds\, g(s)\right) + \alpha a \int_0^\infty ds\, h(s), \end{aligned} \tag{24}$$

where k, g and h are continuous functions, all equal to 1 at 0, satisfying the following differential equations:

$$k'(t) = \begin{cases} a\, k(t) - 1, & 0 \le t < 1 \\ a\{k(t) - k(t-1)\}, & t > 1, \end{cases} \tag{25}$$

$$g'(t) = \begin{cases} a^2 b\, k^2(t) + a\, g(t) - 1, & 0 \le t < 1 \\ a^2 b\{k^2(t) - k^2(t-1)\} + a\{g(t) - g(t-1)\}, & t > 1. \end{cases} \tag{26}$$

and

$$h'(t) = \begin{cases} a^3 c\, k^3(t) + 3a^2 b\, k(t) g(t) + a\, h(t) - 1, & 0 \le t < 1 \\ a^3 c\{k^3(t) - k^3(t-1)\} + 3a^2 b\{k(t)g(t) - k(t-1)g(t-1)\} & \\ +a\{h(t) - h(t-1)\}, & t > 1. \end{cases} \tag{27}$$

Suppose that k, g and h can be written as a power series in a. Then we can evaluate via (24) - (27), in theory, recursively the coefficients of the power series (in a) of $\boldsymbol{E}|Q|^3$. With the use of symbolic manipulation packages the extensive calculations can actually be carried out. The following solution is suggested:

$$\begin{aligned} \boldsymbol{E}|Q|^3 &= \left(\frac{\alpha a}{2(1-a)}\right)^3 + \frac{3}{2}\frac{\alpha^2 a^2}{1-a}\left\{\frac{1}{2(1-a)} + \frac{a^2 b}{12}\left(\frac{4}{1-a} + \frac{3}{(1-a)^2}\right)\right\} \\ &+ \alpha a \left\{\frac{1}{2(1-a)} + \frac{a^2 b}{4}\left(\frac{4}{1-a} + \frac{3}{(1-a)^2}\right) + \frac{b^2 a^4 (11 - a - a^2)}{36(1-a)^3}\right. \\ &+ \left.\frac{c a^3 (9 - 2a - a^2)}{(1-a)^2} + \eta(a) a^4 \left(b^2 - c\right)\right\}, \end{aligned} \tag{28}$$

where η satisfies the expansion

$$\begin{aligned} \eta(a) &= -\frac{1}{180}a - \frac{1}{180}a^2 + \frac{1}{2520}a^3 + \frac{17}{60480}a^4 - \frac{17}{75600}a^5 - \frac{127}{1330560}a^6 \\ &+ \frac{223}{6652800}a^7 + \frac{30047}{1037836800}a^8 - \frac{293}{330220800}a^9 - \frac{1361287}{217945728000}a^{10} \\ &- \frac{226927}{232475443200}a^{11} + \frac{14161487}{11856247603200}a^{12} + \frac{123995383}{266765571072000}a^{13} + \cdots \end{aligned}$$

This indicates that the third moment of $|Q_n^*|$ indeed converges to that of Q^*. It is an open question how the complete power series of η can be found or "guessed".

References

[1] Athreya, K.B. and Ney, P.E. (1972). *Branching Processes.* Springer-Verlag, Berlin.

[2] Boxma, O.J. and Takagi, H. (editors) (1992). Special issue on polling systems. *Queueing Systems* **11** (1,2).

[3] Coffman, Jr. E.G. and Gilbert, E.N. (1986). A continuous polling system with constant service times. *IEEE Trans. Inform. Theory* **32** 584-591.

[4] Coffman, Jr., E.G., Puhalskii, A.A. and Reiman, M.I. (1995). Polling systems with zero switchover times: a heavy-traffic averaging principle. *Ann. Appl. Prob.* To appear.

[5] Daley, D.J., Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes.* Springer, New York.

[6] Feller, W. (1970). *An Introduction to Probability Theory and Its Applications II* (2nd Ed). Wiley, New York.

[7] Fuhrmann, S.W. and Cooper, R.B. (1985). Application of decomposition principle in M/G/1 vacation model to two continuum cyclic queueing models - especially token-ring LANs. *AT & T Techn. J.* **64** 1091-1098.

[8] Jagers, P. (1975). *Branching Processes with Biological Applications.* Wiley, London.

[9] Kingman, J.F.C. (1962). On Queues in Heavy Traffic. *J. Roy. Stat. Soc.* **B24** 383-392.

[10] Kingman, J.F.C. (1982). Queue Disciplines in Heavy Traffic. *Math. of Oper. Res.* **7** 262-271.

[11] Kroese, D.P. and Schmidt, V. (1992). A continuous polling system with general service times. *Ann. Appl. Prob.* **2** 906-927.

[12] Kroese, D.P. and Schmidt, V. (1994). Single-server queues with spatially distributed arrivals. *Queueing Systems.* **17** 317-345.

[13] Resing, J.A.C. (1993). Polling systems and multiple branching processes. *Queueing Systems* **13** 409-426.

[14] Szczotka, W. (1990) Exponential approximation of waiting time and queue size for queues in heavy traffic. *Adv. Appl. Prob.* **22** 230-240.

[15] Takagi, H. (1986). *Analysis of Polling Systems.* MIT Press, Cambridge (MA).

[16] van der Mei, R. (1995). *Polling Systems and the Power Series Algorithm.* Ph.D. Thesis, Tilburg University, The Netherlands.

[17] Whitt, W. (1974). Heavy traffic limit theory for queues: a survey. In: *Mathematical Methods in Queueing Theory* (A.B. Clarke ed.). Lecture Notes in Economics and Mathematical Systems 98, 307-350. Springer-Verlag, Berlin.