Light-Traffic Analysis for Queues with Spatially Distributed Arrivals

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

Volker Schmidt
Department of Stochastics
University of Ulm
Helmholtzstrasse 18, 89069 Ulm
Germany

September 18, 2013

Abstract

We consider the following continuous polling system: Customers arrive according to a homogeneous Poisson process (or a more general stationary point process) and wait on a circle in order to be served by a single server. The server is "greedy", in the sense that he always moves (with constant speed) towards the nearest customer. The customers are served according to an arbitrary service time distribution, in the order in which they are encountered by the server. First-order and second-order Taylor-expansions are found for the expected configuration of customers, for the mean queue length, and for expectation and distribution function of the workload. It is shown that under light traffic conditions the greedy server works more efficient than the cyclically polling server.

Keywords: queueing theory, continuous polling system, greedy server, Poisson arrivals, light-traffic derivative, Taylor-expansion, queue length, workload, general stationary input.

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

1 Introduction

Queueing systems in which customers arrive at a "continuum" of stations – rather than at a finite number of stations – have been studied in recent years as convenient descriptions of transportation systems, machine repair systems, Local Area Networks, computer disk systems, etc., cf. [2], [4], [8], [9], [10], [18], [22], [23], [24], [27]. Often, these continuous models are more transparent and easier to analyze than their discrete analogs, where the customers are assumed to arrive at a finite number of fixed service stations (see, e.g., [34]).

The best studied "continuum queueing model" is the *scanning* or *(cyclically) polling server* model. Here customers arrive (usually according to a Poisson process) on a closed curve (usually a circle) C in the plane. A single server travels at constant speed on C, according to a fixed route which does *not* depend on the actual configuration of customers on C (cf. [8], [10], [18], [22]). Whenever the server encounters a customer he stops and serves this customer according to a fixed service time distribution. This model is the continuous counterpart of discrete polling systems where the server visits consecutive stations according to a (deterministic) "polling table". Recently, *random* polling tables have been investigated as well; see [1], [19], [20], for Bernoulli polling, and [6], [7], [31] for more general Markov polling. The *Brownian server* model, considered in [24], where the server's movement on C is governed by a Brownian motion with drift, can be seen as an approximative continuous model for a certain class of discrete Markov polling systems with many stations. Alternatively, the Brownian server can be interpreted as a cyclically polling server whose movement is disturbed from the outside. In the rest of the paper we will refer to the cyclically polling server simply as "the polling server".

A common feature of the polling and Brownian server model is that the movement of the server is not influenced by the actual "state" of the system, i.e. the actual configuration of customers on C. This is not the case for the so-called *greedy server* model. The greedy server always walks (at constant speed) towards the *nearest* customer on C. A newly arriving customer can thus change the direction of the server's movement. This makes the performance analysis of the greedy server much more difficult than that of the polling and Brownian server. However for all these models an interesting decomposition principle holds (under Poisson arrivals). It says that certain performance characteristics (e.g. the stationary mean queue length and the stationary mean workload) can be represented as the sum of two corresponding conditional mean values given that the server is walking (i.e. not serving) and given that all customers arrive at one and the same point (i.e. no walking times appear), respectively. This means in particular that the second

summand of this sum is a performance characteristic of the "usual" M/G/1 queue and, consequently, well-known. Thus, in case when also the first summand can be determined, a performance analysis of the continuous polling system as a whole is possible (see, e.g., [18], [24]). Note that analogous decomposition results hold not only for mean values, but for the corresponding distributions as well (see [7], [13], [14], [17], [26]).

For the greedy server it seems not possible to determine the first summand of the stochastic decomposition result analytically. For this reason, the performance analysis of the greedy server is very difficult. Moreover, even the question seems to be open whether the usual condition "traffic intensity less than one" is sufficient for stability of the greedy-server queue with spatially distributed Poisson arrivals (cf. [24]).

A natural question to ask is whether the greedy server is always better than the polling server. Intuitively this should be the case when dealing with light-traffic conditions, i.e. for small enough arrival intensity. A comparative (simulation) study of the greedy and polling server was discussed in [9], for the case that C is a circle (or a line interval), the arrival epochs form a homogeneous Poisson process with intensity a, the speed of a server is α^{-1}, and the service times are deterministic with duration e_1. Remarkably, the simulation study indicates that, for a wide range of parameter vectors (α^{-1}, e_1, a), the polling and greedy servers are roughly equally effective. Indeed, the polling server is sometimes better. However, as expected, the greedy server gives substantial improvements in case of light traffic (i.e. small a).

In the present paper we show that analytic techniques can be fruitfully used to examine queues with spatially distributed arrivals. The greedy server, who has proved to be notoriously elusive, can in fact be tackled via light-traffic arguments. We show how first- (and second-) order Taylor-expansions (with respect to a at $a = 0$) for several performance measures can be derived, in particular for the greedy server and for the polling server with non-Poisson arrivals. These expansions could be used (possibly in combination with heavy-traffic results) to construct approximations for several performance measures under "moderate" traffic conditions. A paradoxic result is found concerning the *configuration* of waiting customers on C for the greedy server model. Moreover, it is shown that under quite general assumptions the greedy server is "better" than the polling server in light traffic. That is, we show that with respect to queue length, waiting time and workload, the greedy server works more efficient in light traffic than the polling server.

2 The Polling and Greedy Server Models

2.1 Description of the Models

Consider a circle C with circumference 1. We assume, when not noted otherwise, that customers arrive according to a Poisson process with intensity a; $0 < a < \infty$. In Section 4 we admit however that the arrival epochs form a general stationary ergodic point process with finite intensity. Upon arrival the customers choose their positions on C according to a uniform distribution (independently of everything else), and wait there for service. A single server travels on the circle until he meets a customer. He then stops and serves this customer. We consider two different servers : (a) the *polling server* who moves uni-directionally at constant speed α^{-1} (when not serving; $0 < \alpha^{-1} < \infty$) and (b) the *greedy server* who always walks (at constant speed α^{-1}) towards the *nearest* customer on C. Thus, any newly arriving customer can change the direction of the greedy server's movement, whereas the movement of the (walking) polling server is not influenced by changes of the actual configuration of customers on C. The customers are served in the order in which they are encountered by the server. The consecutive service times are independent of each other (and of everything else) and identically distributed having an arbitrary distribution function F with finite first two moments e_1, e_2. After service completion, the customer is removed from the circle and the server resumes his walk. For both models we assume that there is no customer on C at time $t = 0$.

2.2 Definition of Performance Characteristics

An important performance characteristic for both models is the stationary *mean number of customers* on the circle. Another, more detailed performance characteristic, unique to queueing systems with a spatially "continuous" waiting room, is the stationary *expected configuration of waiting customers* given that the server is walking (i.e. not serving). We will specify first in which mathematical framework these quantities can be precisely defined.

Note that, in steady state, the actual position of the server is not very important to the analysis of the model. It is the configuration of customers *relative* to the position of the server that counts. We therefore restate the models as seen from the point of view of the server, in terms of the following particle systems:

Particle System 1a. On the strip $\mathbb{D} = \mathbb{R}_+ \times [0, 1]$ particles are born and

die according to the following rules. Particles are born according to a (two-dimensional) homogeneous Poisson point process on $\mathbb{D}$ with mean measure $a\,\nu$, where ν denotes the Lebesgue measure in $\mathbb{D}$. All particles move at constant speed α^{-1} towards level 0. When a particle hits 0, all particles stop moving during a random amount of time which is independent of everything else, and which has distribution function F, after which the particle dies that hit the zero level, and the movement of the particles is continued.

Particle System 1b. It is defined in the same way as particle system 1a with the exception that particles move in a different way. Namely, when the uppermost particle is further away from 1 than the lower-most one is from 0, all particles move at speed α^{-1} towards 0. On the other hand, when the uppermost particle is closer to level 1 than the lower-most is to level 0, all points move at speed α^{-1} towards 1. Consequently, in the present case, particles die after hitting one of the levels 0 or 1 and after spending there a time with distribution function F.

For both particle systems 1a and 1b, let W_t denote the *random counting measure* on the Borel σ-algebra $\mathcal{B}([0,1])$ whose atoms are the positions of the particles that are "alive" at time $t \geq 0$. Clearly, W_t corresponds to the positions of waiting customers *relative* to the position of the server at time t. As in [22] (see also [23], [24]) it will be convenient to introduce a further kind of particle systems which is obtained from the particle systems 1a and 1b by "deleting" the times when the particles do not move.

Particle Systems 2a, 2b. There are two different kinds of particles, *parents* and *offspring.* Parent particles are born according to a (two-dimensional) homogeneous Poisson point process on $\mathbb{D}$ with mean measure $a\,\nu$. In the systems 2a and 2b, particles move in the same way as in systems 1a and 1b, respectively. When a particle hits 0 or 1, it dies immediately, but, simultaneously, a random number of offspring is produced at that time, the position of each individual child being uniformly distributed on $[0,1]$, independently of each other and of everything else. The number of offspring generated by a single dying particle has the same distribution as the random variable J with generating function

$$\boldsymbol{E}\, z^J = \int_0^\infty \mathrm{e}^{-a(1-z)u} dF(u), \quad 0 \leq z \leq 1. \tag{2.1}$$

Let Q_t denote the random counting measure on $\mathcal{B}([0,1])$ corresponding to the configuration of particles that are "alive" at time $t \geq 0$ in the second kind of particle systems. The measure-valued stochastic process (Q_t) describes the evolution of the configuration of waiting customers on C given that the server is walking (i.e. not serving).

Besides the two objects Q_t and W_t that describe the configuration (and, in particular, the number) of waiting customers at time t, another important performance measure is the *workload* on the circle. We can define the workload in terms of particle systems 1a and 1b in the following way. Suppose at time t, n particles are alive and are not at level 0 or 1. Each of these particles will eventually hit 0 or 1 and will then die after a "service time" $y_1, \ldots, y_n$, respectively. The sum of these times is denoted by Z_t. When one of the n particles is at 0 or 1, Z_t is defined as the sum of the residual lifetime of that particle and the "service times" of the other $n-1$ particles. Notice that the random variable Z_t can be interpreted as the workload on the circle at time t.

Furthermore, we introduce the workload process (R_t), by defining R_t to be the sum of the "service times" of the particles that are alive at time t in particle systems 2a or 2b. Notice that one never actually observes these service times in these particle systems. A proper way to define (R_t) would be via a random time change of (Z_t). Note that R_t has the same distribution as $S_1 + \cdots + S_{|Q_t|}$, where $S_1, S_2, \ldots$ is a sequence of i.i.d random variables with distribution function F, which are independent of $|Q_t|$, the number of particles alive at time t in system 2a or 2b. By some authors, a different notion of workload is considered for queues with vacations which is defined by summing up service times *and* adding the residual vacation time provided that the server is actually on vacation. It is called the *virtual workload* (cf. [14], [25], [26]). For continuous polling systems, an analogous virtual workload process can be defined in the following way. Let $\hat{Z}_t$ denote the amount of time that elapses beginning from time t until all particles die being present at time t in the systems 1a and 1b, respectively, provided that in the meantime no new customers arrive. Thus, $\hat{Z}_t$ is obtained by summing up service times and *all* walk times needed for "serving" the particles being in the system at time t. By $(\hat{R}_t)$ we denote the corresponding workload process in the systems 2a and 2b.

It turns out (see next subsection) that the random measures W_t and Q_t converge in distribution to random measures W and Q, respectively. Moreover, Z_t and R_t (or $\hat{Z}_t$ and $\hat{R}_t$) converge in distribution to random variables Z and R (or $\hat{Z}$ and $\hat{R}$), respectively. Other random variables, notably the sojourn time U and the waiting time V of a random customer will be introduced later. These steady-state performance characteristics, in particular the light-traffic behavior of their expectations and distribution functions, will be the subject of our studies in the present paper.

2.3 Stability Conditions and Representation of Steady-State Performance Characteristics

In the following we will, for a (random) measure M on $\mathcal{B}([0,1])$, abbreviate $M([0,1])$ to $|M|$, and, consequently, $\boldsymbol{E}\, M([0,1])$ to $\boldsymbol{E}\,|M|$.

The quantity $\rho = ae_1$ is usually called the *traffic intensity* of the considered single-server system. Under the condition that

$$\rho < 1, \tag{2.2}$$

it has been shown in [24] that, for the polling server, there exist time-stationary regenerative processes $(\tilde{W}_t)$ and $(\tilde{Q}_t)$ such that the finite-dimensional distributions of the processes $(W_{t+h}; t \geq 0)$ and $(Q_{t+h}; t \geq 0)$ converge in variation to the corresponding finite-dimensional distributions of $(\tilde{W}_t)$ and $(\tilde{Q}_t)$, respectively, as h tends to infinity.

On the other hand, the question is still open whether, under (2.2), an analogous stability theorem holds true also for the greedy server (see the remark at the end of Section 3.2 in [24]). However, under light traffic, stability of the greedy server follows from the following argument. Assume that

$$a(e_1 + \frac{\alpha}{2}) < 1, \tag{2.3}$$

where α is the time which the server needs for walking once around the whole circle (without stopping on the way). Then, using Loynes' scheme (see, e.g., Chapter 2 of [16]), one can show that for the greedy server the same stability theorem holds as for the polling server. Namely, a stable auxiliary M/G/1 queue can be constructed (with mean service time equal to $e_1 + \alpha/2$), which works slower than the original greedy server and for which, in particular, each empty point (i.e. an arrival epoch at which the system is empty) is simultaneously an empty point of the greedy-server system. Because we are interested in the *light-traffic behavior* (i.e. $a \to 0$) of queues with spatially distributed arrivals, the strengthened stability condition (2.3) is not restrictive for our purposes. Moreover, under (2.3), Loynes' scheme for constructing the stationary queueing process $(\tilde{W}_t)$ works also under the assumption that the arrival epochs form a general stationary ergodic point process, for both the polling and greedy servers.

For the rest of the paper we use the notation $W = \tilde{W}_0$ and $Q = \tilde{Q}_0$. From the individual ergodic theorem (see, e.g., Theorem 1.3.12 of [16]) it follows that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t W_s(B)\, ds = \boldsymbol{E}\, W(B) \quad \text{for} \quad B \in \mathcal{B}([0,1]) \tag{2.4}$$

and

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t Q_s(B)\, ds = \boldsymbol{E}\, Q(B) \quad \text{for} \quad B \in \mathcal{B}([0,1]), \tag{2.5}$$

with probability one. Thus, the stationary mean measures $\boldsymbol{E}\, W(\cdot)$ and $\boldsymbol{E}\, Q(\cdot)$ are two important performance characteristics.

The random counting measure W can be interpreted as the stationary configuration of customers on the circle relative to the position of the server, at an *arbitrary point in time* (i.e. not specifying whether the server is actually walking or serving). Similarly, Q can be interpreted as the stationary configuration of waiting customers (relative to the server) *given* that the server is *walking*, i.e. "at a random travelling epoch".

In the same way, using Loynes' scheme, under (2.3) one can show for both the polling and greedy servers that there exists a stationary regenerative process $(\tilde{Z}_t)$ such that the finite-dimensional distributions of the workload processes $(Z_{t+h}; t \geq 0)$ converge in variation to the corresponding finite-dimensional distributions of $(\tilde{Z}_t)$. Moreover, with the notation $Z = \tilde{Z}_0$, we get from the individual ergodic theorem that

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t Z_s\, ds = \boldsymbol{E}\, Z \tag{2.6}$$

with probability one. The random variable Z can be interpreted as the stationary workload on C, at an arbitrary point in time. Similarly, R_t converges in variation, to a random variable R, where R can be interpreted as stationary workload on C, at a random travelling epoch. A completely analogous argument can be used for proving the existence of random variables $\hat{Z}$ and $\hat{R}$, representing the *virtual* workload in the stationary situation at an arbitrary point in time and at a random travelling epoch, respectively. Finally, note that the distribution of R can be represented by

$$R \stackrel{\mathrm{d}}{=} \sum_{k=1}^{|Q|} S_k, \tag{2.7}$$

where $S_1, S_2, \ldots$ is a sequence of i.i.d. non-negative random variables with distribution function F, which are independent of Q.

2.4 Stochastic Decomposition

Assume that the arrival process is Poisson and that (2.3) holds. The following *stochastic decomposition* results show that the distributions of $|W|$ and Z are completely specified by the distributions of $|Q|$ and R, and vice-versa.

Proposition 1 For every $p \geq 0$ we have

$$\boldsymbol{E}\,\mathrm{e}^{-p|W|} = \frac{(1-\rho)(1-\mathrm{e}^{-p})L_F(\beta)}{L_F(\beta)-\mathrm{e}^{-p}}\,\boldsymbol{E}\,\mathrm{e}^{-p|Q|} \tag{2.8}$$

and

$$\boldsymbol{E}\,\mathrm{e}^{-pZ} = \frac{(1-\rho)p}{p-a+aL_F(p)}\,\boldsymbol{E}\,\mathrm{e}^{-pR}, \tag{2.9}$$

where $L_F(s) = \int_0^\infty \mathrm{e}^{-st}dF(t)$ and $\beta = a(1-\mathrm{e}^{-p})$.

For a *proof* of (2.8) we refer to [17], for that of (2.9) to [7], see also [13], [26].

Note that the first factor at the right-hand side of (2.8) and (2.9) is equal to the Laplace-Stieltjes transform of queue length and workload, respectively, in the "usual" M/G/1 queue. Unfortunately it seems that, for the virtual workload $\hat{Z}$, an analogous decomposition formula does not hold. At first glance, this might be surprising. Namely, for this kind of workload, a decomposition result similar to (2.9) has been obtained in Theorem 3.3 of [26] for quite general server vacations. Note however that, in continuous polling systems, the increment of the virtual work load process $(\hat{Z}_t)$ at an arrival epoch is *not* independent of the value of $(\hat{Z}_t)$ at this arrival epoch and, consequently, the PASTA argument used in [26] does not work in our case.

Corollary 1 In particular, it holds that

$$\boldsymbol{E}\,|W| = \rho + \frac{a^2e_2}{2(1-\rho)} + \boldsymbol{E}\,|Q| \tag{2.10}$$

and

$$\boldsymbol{E}\,Z = \frac{ae_2}{2(1-\rho)} + \boldsymbol{E}\,R = \frac{ae_2}{2(1-\rho)} + e_1\,\boldsymbol{E}\,|Q|, \tag{2.11}$$

where we have used (2.7) and Wald's lemma in the second equation of (2.11).

It turns out that, sometimes, the random variable $|Q|$ (or the random measure Q) is easier to analyze than $|W|$ (or W). In fact, for the *polling server*, the exact form of the Laplace-Stieltjes transform of $|Q|$ is known (cf. Theorem 4.1 of [22]). Moreover, even when the distribution of $|Q|$ is hard to get – as is the case in the Brownian server model – closed formulas for the mean measure $\boldsymbol{E}\,Q(\cdot)$ are usually much easier to find. The performance characteristics $\boldsymbol{E}\,|W|$ and $\boldsymbol{E}\,Z$ then follow from (2.10) and (2.11).

Remark 1 For the polling server, the mean measure of Q is given by (see [22] or [18]):

$$\boldsymbol{E}\, Q(dx) = \frac{a\alpha}{(1-\rho)}\,(1-x)\,dx, \quad x \in [0,1]. \tag{2.12}$$

and hence,

$$\boldsymbol{E}\,|W| = \rho + \frac{a^2 e_2 + a\alpha}{2(1-\rho)}. \tag{2.13}$$

and

$$\boldsymbol{E}\, Z = \frac{a e_2 + a\alpha e_1}{2(1-\rho)}. \tag{2.14}$$

Another performance characteristic, the stationary mean sojourn time of a "random customer" $\boldsymbol{E}_0 S$, follows from Little's formula (see, e.g., Theorem 4.2.1 of [16]):

$$\boldsymbol{E}\,|W| = a\,\boldsymbol{E}_0 S. \tag{2.15}$$

Here the expectation $\boldsymbol{E}_0$ is taken with respect to the Palm distribution of arrival epochs. Notice that (2.14) also follows from (2.13), (2.15) and Brumelle's formula (see, e.g., Corollary 4.2.2 in [16]):

$$\boldsymbol{E}\, Z = \rho(\,\boldsymbol{E}_0 V + \frac{e_2}{2e_1}), \tag{2.16}$$

where $\boldsymbol{E}_0 V = \boldsymbol{E}_0 S - e_1$ is the stationary mean waiting time of a random customer.

Note that Little's formula (2.15) and Brumelle's formula (2.16) hold for both the polling and the greedy server model. Moreover, for both models these formulas remain true when the arrival epochs form an arbitrary stationary ergodic point process with finite intensity such that (2.3) is fulfilled (see Section 4.2 of [16]). This will be used in Section 4 for getting light-traffic derivatives for $\boldsymbol{E}\,|W|$, $\boldsymbol{E}\, Z$ and $\boldsymbol{E}\,\hat{Z}$ for the polling and the greedy server systems with general input.

For the greedy server, no analytic formulas for probabilistic characteristics of $|Q|$ are known, not even for the expectation $\boldsymbol{E}\,|Q|$ in the Poisson arrival case. This makes it impossible to give, in this case, closed formulas for the performance characteristics $\boldsymbol{E}\,|W|$, $\boldsymbol{E}_0 V$ and $\boldsymbol{E}\, Z$. However, it is possible to derive asymptotic results for these performance characteristics and for the mean measure of Q, for small a.

3 Light-Traffic Results for the Greedy Server Model

In Section 2.4, several performance characteristics have been given for the polling model, but no such results exist for the greedy server model. However, by using light-traffic analysis we can derive second-order Taylor-expansions for several performance characteristics, e.g. for $\boldsymbol{E}\,|W|$ and $\boldsymbol{E}\,Z$. These Taylor-expansions could be used to obtain approximations to the true performance characteristics. They show that the greedy server is indeed considerably better than the polling one, when dealing with light traffic.

Moreover, we obtain an expansion for the expected *configuration* of customers, given that the server is traveling. For the polling server, the expected density of waiting customers (given that the server is not busy) is given in (2.12). This is a quite intuitive result: The density is highest in front of the server, and decreases linearly with the distance in front of the server. For the greedy server, one would (just as intuitively) expect that the density of waiting customers (given that the server is not busy) would be maximal at locations opposite to the server, as in the drunken server case (cf. [23]). But this is not true, at least not in light traffic. On the contrary, the density is minimal at a distance $1/2$ of the server. We can see this in the following theorem, where a (second-order) Taylor-expansion (at $a = 0$) for $\boldsymbol{E}\,Q([0,x])$ is given, for all $x \in [0,1]$, viewing this quantity as a function of a.

Theorem 1 The following expansion holds for all $x \in [0,1]$:

$$\boldsymbol{E}\,Q([0,x]) = \int_0^x m_Q(u)\,du + O(a^3) \tag{3.1}$$

as $a \to 0$, where

$$m_Q(u) = a\alpha(1+\rho)\left(\frac{1}{2} - u\right) + a^2\alpha^2\left(\frac{u^2}{2} - \frac{2u^3}{3}\right) \tag{3.2}$$

for $u \in [0,1/2]$, and $m_Q(u) = m_Q(1-u)$ for $u \in [\frac{1}{2}, 1]$.

The *proof* of the theorem is given at the end of this section.

Notice that Theorem 1 suggests that m_Q, defined in (3.2), is the second-order expansion of the density of $\boldsymbol{E}\,Q(\cdot)$ in light traffic. But, formally, we have not proved the existence of such a density. However, from a pragmatic point we tacitly assume that such a density exists. Note that m_Q is different from what one would expect at first sight. Since the greedy server always moves towards the nearest customer, one would expect that "opposite" to the server (at the opposite side on the circle) would be the most customers

waiting. However, (3.2) indicates that in light traffic the *contrary* is true: the (first-order expansion for the expected) density of customers at u is *least* for $u = 1/2$. An explanation for this is that the density should be viewed as a mixture of two densities, one *given* that the server moves in a clockwise fashion, and the other given that the server moves in a counter-clockwise fashion. Each such density resembles the density in the polling server model, that is a (in first order linearly) decreasing density.

The following corollary follows immediately from Theorem 1, (2.10) and (2.11).

Corollary 2 The second-order expansion for the mean number of customers is:

$$\boldsymbol{E}\,|W| = a\left(e_1 + \frac{\alpha}{4}\right) + a^2\left(\frac{e_2}{2} + \frac{\alpha e_1}{4} + \frac{\alpha^2}{48}\right) + O(a^3). \tag{3.3}$$

And the second-order expansion for the mean workload is given by:

$$\boldsymbol{E}\,Z = a\left(\frac{e_2}{2} + \frac{e_1\alpha}{4}\right) + a^2\left(\frac{e_1 e_2}{2} + \frac{\alpha e_1^2}{4} + \frac{\alpha^2 e_1}{48}\right) + O(a^3). \tag{3.4}$$

Notice that (3.3) and (3.4), when compared to (2.13) and (2.14), show that (in light traffic) the greedy server is considerably more effective than the polling server.

Remark 2 From the (second-order) Taylor-expansions we can derive approximations of the performance characteristics by simply dropping the $O(a^3)$ term. In order to see how good the approximations are, simulation studies should be carried out. This will be done in a separate paper. If also *heavy traffic* results would be available, these Taylor-expansions could be used to get via an appropriate interpolation, approximation formulas for an arbitrary positive arrival rate a, see [15], [28], [36]. The behavior of the greedy server in heavy traffic is a very interesting (open) problem. All that is known (through simulation studies) is that the greedy server behaves in heavy traffic as a polling server, that is, during a busy period the greedy server will either travel almost always in a positive direction (with probability 1/2) or almost always in a negative direction.

In our search for light-traffic results, we are led by [3], [5] and [30] (see also [29], [32], [33], [35], [36]). In particular, we will use an extension of Corollary 5.2 in [5] to the case of an independently marked Poisson process. Since we are now concerned with stationary characteristics, we assume, in view of Section 2, that the arrival times $\{T_n\}$ form a Poisson process on the whole

real line $I\!R$ with intensity a, where we make the convention

$$\ldots < T_{-1} < T_0 < 0 \leq T_1 < T_2 < \ldots$$

Assume also that (2.3) holds, i.e.

$$a(e_1 + \frac{\alpha}{2}) < 1. \tag{3.5}$$

Observe that we are in the same framework as [30], that is, we are dealing with a marked Poisson process $\{(T_n, Z_n)\}$ on $I\!R$, with markings $Z_n = (X_n, S_n)$, where X_n denotes the position (relative to the position of the server at time T_n) and S_n the service time of the customer arriving at T_n. These markings are are i.i.d. and independent of the Poisson process $\{T_n\}$.

Next, consider other marked point processes with the *same* i.i.d. markings as $\{(T_n, Z_n)\}$, independent of the (non-marked) point process. For such processes, let W_t be the random measure of waiting customers at time t, $t \in I\!R$. And, for $t \in I\!R$, let

$$L_t = W_t([z, z + dz])I_{\{W_t(\{0,1\})=0\}}, \tag{3.6}$$

where $z \in (0, 1)$, dz is some small strictly positive number, and I_B denotes the indicator of the set B. That is, L_t denotes the number of customers at time t waiting in the interval $[z, z + dz]$, provided that the server is not busy, or else $L_t = 0$.

The idea behind the light-traffic analysis is to regard $\boldsymbol{E}\, L_0$, for the marked Poisson process $\{(T_n, Z_n)\}$, as a function f of the arrival intensity a, and to obtain a (Taylor-) expansion of f in a at 0. This is accomplished by considering the behaviour of L_t for input processes that consist of only a few arrivals. For example, suppose that there is only one arrival on $I\!R$, at time s. Define $G(s, x) = L_x$ and let $g(s, x) = \boldsymbol{E}\, G(s, x)$. Analogously, for the case that the arrival process consists only of two arrivals, at s and t, we define $H(s, t, x) = L_x$ and $h(s, t, x) = \boldsymbol{E}\, H(s, t, x)$. We will show in Theorem 2 (which is an extension of Corollary 5.2 in [5]) that the first and second (right-hand) derivatives $f^{(1)}(0)$ and $f^{(2)}(0)$ of f at $a = 0$ exist, under the condition that (3.5) holds and that the second moment e_2 of service times is finite. Moreover, Corollary 5.2 in [5] provides a way to calculate these derivatives via the functions h and g, defined above (see also Theorem 2 of [30], where an additional "admissibility" condition is assumed). Before we state Theorem 2 we give some preliminaries:

Let $\{(T_n, Z_n)\}$ be an independently marked stationary Poisson process with the mark space $K = [0, 1] \times I\!R_+$. By Ω_K we denote the set of realizations

of $\{(T_n, Z_n)\}$, i.e. Ω_K is the set of (locally finite) counting measures ω on $I\!R \times K$ with

$$\omega([a,b] \times K) < \infty,$$

for every bounded interval $[a,b]$. Let ψ be a real-valued (measurable) functional defined on Ω_K. For every $t \in I\!R$, let the restriction $\omega|_t$ of $\omega \in \Omega_K$ be defined by

$$\omega|_t(B \times C) = \omega(B \cap (-\infty, t) \times C).$$

Furthermore, for any $t \in I\!R$ and $z \in K$, let

$$\psi_{(t,z)}(\omega) = \psi(\omega|_t + \delta_{(t,z)}) - \psi(\omega|_t)$$

and, for any $t_1, \ldots, t_n \in I\!R$ and $z_1, \ldots, z_n \in K$, let

$$\psi_{(t_1,z_1),\ldots,(t_n,z_n)}(\omega) = (\ldots(\psi_{(t_1,z_1)})_{(t_2,z_2)}\ldots)_{(t_n,z_n)}(\omega).$$

Note that the functional $\psi_{(t_1,z_1),\ldots,(t_n,z_n)}$ can be written in the form

$$\psi_{(t_1,z_1),\ldots,(t_n,z_n)} = \begin{cases} \sum_{j=0}^{n}(-1)^{n-j} \sum_{\pi \in \{\binom{n}{j}\}} \psi(\omega|_{t_n} + \sum_{i \in \pi} \delta_{(t_i,z_i)}) & \text{for } t_n < \ldots < t_1 \\ 0 & \text{otherwise,} \end{cases} \tag{3.7}$$

where $\{\binom{n}{j}\}$ denotes the collection of all those subsets of $\{1, \ldots, n\}$ containing j elements.

Following [5], we call the functional ψ *continuous at infinities* if

$$\lim_{t \to -\infty} \psi(\omega|_t) = \psi(\mathbf{0})$$

and

$$\lim_{t \to \infty} \psi(\omega|_t) = \psi(\omega),$$

for every $\omega \in \Omega_K$, where $\mathbf{0}$ denotes the zero measure with $\mathbf{0}(I\!R \times K) = 0$.

The following result is a straightforward extension of Corollary 5.2 in [5], where the non-marked case has been considered.

Theorem 2 Let P_a denote the distribution of the independently marked Poisson process $\{(T_n, Z_n)\}$ with intensity a, and let $\tilde{F}$ denote the distribution of the random variables Z_n. If the functional ψ is continuous at infinities, if

$$\int_{(I\!R \times K)^i} \int_{\Omega_K} |\psi_{(t_1,z_1),\ldots,(t_i,z_i)}(\omega)| P_a(d\omega) dt_1 \tilde{F}(dz_1) \ldots dt_i \tilde{F}(dz_i) < \infty \tag{3.8}$$

for all $i = 1, \ldots, n+1$, and if

$$\limsup_{a \to 0} \int_{(I\!R \times K)^{n+1}} \int_{\Omega_K} \psi_{(t_1,z_1),\ldots,(t_{n+1},z_{n+1})}(\omega) P_a(d\omega) \\ dt_1 \tilde{F}(dz_1) \ldots dt_{n+1} \tilde{F}(dz_{n+1}) = 0, \tag{3.9}$$

then

$$\boldsymbol{E}\psi(\{(T_n, Z_n)\}) = \psi(\mathbf{0}) + \sum_{i=1}^{n} a^i \int_{(I\!R \times K)^i} \psi_{(t_1,z_1),\ldots,(t_i,z_i)}(\mathbf{0}) \\ dt_1 \tilde{F}(dz_1) \ldots dt_i \tilde{F}(dz_i) + O(a^{n+1}). \tag{3.10}$$

Proof of Theorem 1. We apply Theorem 2 to the functional $\psi = L_0$ defined in (3.6). It is easy to see that this functional is continuous at infinities. Furthermore, from (3.7) it follows that condition (3.8) is fulfilled if $\boldsymbol{E}\, L_0 < \infty$. However, $\boldsymbol{E}\, L_0 \leq \boldsymbol{E}\, |W|$. Because we assumed that $e_2 < \infty$, the expectation $\boldsymbol{E}\, |W|$ is finite for all a satisfying (3.5). By similar arguments we get that also (3.9) is fulfilled. Thus, the Taylor expansion (3.10) holds for $\psi = L_0$. Specifically:

$$f^{(1)}(0) = \int_{-\infty}^{\infty} g(s, 0)\, ds \tag{3.11}$$

and

$$f^{(2)}(0) = 2 \int_{-\infty}^{\infty} \int_{-\infty}^{s} \left(h(s,t,0) - g(s,0) - g(t,0) \right)\, dt\, ds. \tag{3.12}$$

Trivially, $f(0) = 0$. Moreover, it is easy to see that

$$\boldsymbol{E}\, Q([z, z+dz]) = f(a)/(1-\rho), \tag{3.13}$$

for $z \in (0,1)$. In order to prove (3.1), it remains therefore to show that for all $z \in (0, 1/2]$

$$f^{(1)}(0) = \alpha \left(\frac{1}{2} - z \right)\, dz + o(dz) \tag{3.14}$$

and

$$f^{(2)}(0) = \alpha^2 \left(z^2 - \frac{4z^3}{3} \right)\, dz + o(dz), \tag{3.15}$$

as $dz \to 0$. The result for $z \in [1/2, 1)$ follows then by symmetry.

Instead of evaluating the integrals (3.11) and (3.12) we will use a convenient transformation which reduces the calculations to "integration over paths". Notice that $h(s,t,0) = h(0,t-s,-s)$, $g(s,0) = g(0,-s)$ and $g(t,0) = g(0,-t)$. Substitute this into (3.11) and (3.12). Now perform the transformation $(s,t) \to (u(s,t), v(s,t))$, where $u(s,t) = t-s$ and $v(s,t) = -s$. Finally use the fact that for $u \geq 0$, $h(0,u,v) = g(0,v) = 0$, for all $v < 0$. This leads to the following results

$$f^{(1)}(0) = \int_0^\infty g(0,v)\, dv \tag{3.16}$$

and

$$f^{(2)}(0) = 2\int_0^\infty \int_0^\infty (h(0,u,v) - 2g(0,v))\, dv\, du. \tag{3.17}$$

From (3.16) it is easy to see that (3.14) holds. Next, we determine, for fixed u, the integral $\int_0^\infty h(0,u,v)\, dv$. Consider thus the situation that there are only two arrivals, one customer arriving at time 0, at a distance X (say) from the server and another arriving at time u, at a distance Y (say) from the server. We may assume that X and Y are i.i.d. r.v.'s, uniformly distributed on $[0,1]$. If $u \geq \alpha/2$, we are certain that the second customer does not arrive before the server has reached the first customer. Consequently,

$$\int_0^\infty h(0,u,v)\, dv = 2\int_0^\infty g(0,v)\, dv, \text{ for } u \geq \alpha/2. \tag{3.18}$$

Let $\boldsymbol{E}^x$ denote the conditional expectation given "$X = x$", and define

$$k(x,u,z,dz) := \boldsymbol{E}^x \int_0^\infty H(0,u,v)\, dv, \tag{3.19}$$

for $(x,u) \in [0,1] \times [0,\alpha/2]$. Below, this quantity will be calculated (up to $o(dz)$). Notice, that by the symmetry of the model,

$$k(x,u,z,dz) = k(1-x,u,1-z,dz) + o(dz). \tag{3.20}$$

Hence combination of (3.14)–(3.20) shows that (3.15) is proved if we can show that

$$2\int_0^{\alpha/2} \int_0^{1/2} [\, k(x,u,z,dz) + k(x,u,1-z,dz)\,]\, dx\, du$$

$$= \alpha^2 \left(1 - 2z + z^2 - \frac{4z^3}{3}\right) dz + o(dz), \tag{3.21}$$

for all $z \in (0,1)$, as $dz \to 0$.

For $u > \alpha x$ and $x < 1/2$ the server reaches the first customer (who arrived at 0) before the arrival of the second customer. It is easy to see that for this case, for $z \in [0,1]$ and $dz \to 0$,

$$k(x,u,z,dz) = \alpha\, dz \left(I_{[z,1/2]}(x)\, I_{[0,1/2]}(z) + |1/2 - z| \right) + o(dz). \tag{3.22}$$

The more difficult case arises when the second customer arrives before the first customer has been reached, that is when $u \le \alpha x$ (again, we only consider the case $x < 1/2$). It is now possible that the server reverses his direction. Before time u there is only one customer in the system travelling with constant speed α^{-1} towards 0 (from the perspective of the server), so

$$\boldsymbol{E}^{\,x} \int_0^u H(0,u,v)\, dv = \alpha\, dz\, I_{[x-\alpha^{-1}u,x]}(z) + o(dz). \tag{3.23}$$

At time u there are two particles in the system. In connection with this, consider the following deterministic particle system:

Particle System 3. Suppose at time 0 two particles are born, at positions $x \in [0,1/2]$ and $y \in [0,1]$. These particle moves towards 0 or 1 in the "greedy server way" (as in the particle systems 1 and 2). They die instantaneously when they hit 0 or 1 without producing offspring. All possible particle movements are depicted in Figure 1. For $(x,y,z) \in [0,1/2] \times [0,1] \times [0,1]$, let $\lambda(x,y,z)$ be the number of times that a particle trajectory crosses level z (see figure 1). And let

$$r(x,z) = \int_0^1 \lambda(x,y,z)\, dy. \tag{3.24}$$

Notice that the value of $r(x,z)$ does not depend on α. A moment of reflexion will show that

$$\boldsymbol{E}^{\,x} \int_u^\infty H(0,u,v)\, dv = \alpha\, dz\, r(x - \alpha^{-1}u, z) + o(dz). \tag{3.25}$$

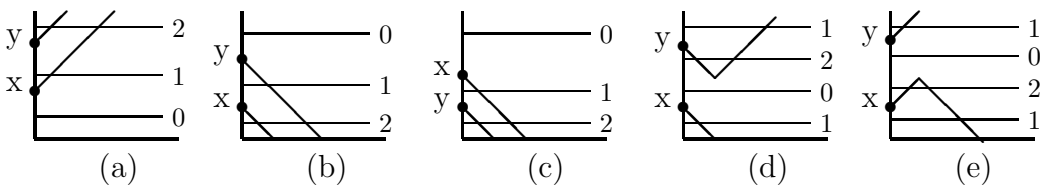
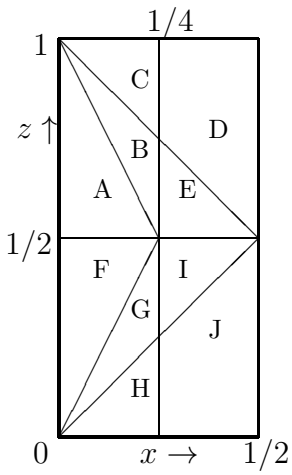


Figure 1. *Possible movements of two particles starting at* $x \in [0,1/2]$ *and* y. $\lambda(x,y,z)$ *is given at the end of the horizontal line-segment at hight* z.

It remains therefore to calculate $r(x,z)$ for $(x,z) \in [0,1/2] \times [0,1]$. These calculations form the most involved part of the proof, but they are straightforward. In Figure 2 a complete specification of r is given.



A : $-1/2 + z + x$
B : $3/2 - z - 3x$
C : $-1/2 + z - x$
D : $-3/2 + z + 3x$
E : $1/2 - z + x$
F : $1/2 - z + x$
G : $1/2 - 3z + 5x$
H : $3/2 - z + x$
I : $3/2 - 3z + x$
J : $5/2 - z - 3x$

Figure 2. *The values for* $r(x,z)$.

The reader may check that from (3.23), (3.25) and Figure 2 it follows that

$$\int_0^{1/2} \int_0^{\alpha x} k(x,u,z,dz)\, du\, dx =$$

$$\begin{aligned} &\alpha^2 dz\ (19 - 12z - 24z^2 - 40z^3)\,/96 + o(dz) \qquad \text{if } z \in [0,1/2] \\ &\alpha^2 dz\ (-7 + 36z - 48z^2 + 24z^3)\,/96 + o(dz) \qquad \text{if } z \in [1/2,1] \end{aligned} \tag{3.26}$$

On the other hand, from (3.22) it follows that for all $z \in [0,1]$

$$\int_0^{1/2} \int_{\alpha x}^{\alpha/2} k(x,u,z,dz)\, du\, dx =$$

$$\alpha^2\, dz \left(\frac{(1/2 - z)^2}{4} + \frac{|1/2 - z|}{8} \right) + o(dz). \tag{3.27}$$

The combination of (3.20), (3.26) and (3.27) now yields the desired result (3.21), which completes the proof of Theorem 1. □

Remark 3 Note that, using the same approach, it should (in principle) be possible to derive higher-order expansions for $\boldsymbol{E}\, Q(\cdot)$. It is also possible to

derive the expansions for $\boldsymbol{E}\,|W|$ and $\boldsymbol{E}\,Z$ directly, indeed the analysis is somewhat simpler. The Taylor-expansion for $\boldsymbol{E}\,W(\cdot)$ can also be computed via the same method.

Remark 4 The non-marked version of Theorem 2 has been obtained in [5] as a corollary of a much more general expansion for functionals of arbitrary stationary (non-marked) point processes, where the notions of higher-order Campbell measures and Palm distributions of such processes have been used. This approach seems to be very promissing for deriving (higher-order) light-traffic approximations for queues with a (non-Poissonian) arrival process possessing some special dependence structure which still allows to calculate limits of higher-order Palm distributions and factorial moment measures when the arrival intensity a tends to zero. In [3], conditions different from those in [5] and [30] have been given for the validity of analogous formulas for light-traffic derivatives, where in [3] mostly the case has been considered that the input is a general stationary marked point process. The crucial step of the approach given in [3] is to show that a certain functional g is non-negative almost surely (see condition (i) of Theorem $1'$ in [3]) and, then, to use Campbell's formula for stationary marked point processes. Unfortunately, for the work loads Z and $\hat{Z}$ in the greedy-server model (and, even more, for the queue length $|W|$) it is difficult to check the conditions given in [3] because their non-negativity condition (with respect to the functionals g corresponding to these queueing characteristics) seems not to hold for the greedy-server model. However, we show in Section 4 that (first-order) light-traffic results for the greedy server can indeed be derived under mild conditions on the input process.

4 Light-Traffic Derivatives for General Stationary Ergodic Arrival Processes

In this section we show how standard formulas of queueing and point process theories can be used in order to get, in an elementary way, light-traffic derivatives for stationary characteristics of (total) queue length and workload of the polling server (PS) and of the greedy server (GS) with general input considering these characteristics as functions of the arrival intensity a. For getting first-order light-traffic derivatives in the greedy-server model (with not necessarily Poisson arrival process), no additional condition on the service time distribution is needed, where we do not use Campbell's formula in its general form. It turns out that two well-known specifications of the general Campbell formula, i.e. Little's and Brumelle's formulas, are more appropriate tools for the purpose of Section 4.1.

The arrival epochs are assumed to form a quite general stationary ergodic (neither necessarily Poisson nor recurrent) point process satisfying, however, the following conditions. For each $a \in (0, (e_1 + \frac{\alpha}{2})^{-1})$, consider a stationary ergodic point process of arrival epochs $(T_n^{(a)})$ with intensity a. Let A_x denote the event that the customer arriving at zero, under the Palm distribution $\boldsymbol{P}_0$ of arrival epochs, finds no further customers in the system and that the next inter-arrival time is greater than x.

Assumption We assume that

$$\lim_{a \to 0} \boldsymbol{P}_0(A_\alpha) = 1 \tag{4.1}$$

and that there is a random variable $\tilde{V}$ such that

$$\boldsymbol{E}_0 \tilde{V} < \infty \quad \text{and} \quad V \leq \tilde{V} \quad \text{for all sufficiently small } a. \tag{4.2}$$

Note that the conditions (4.1) and (4.2) are fulfilled, e.g., in the case when (i) the arrival processes $(T_n^{(a)})$ are obtained by dilation of a fixed point process (T_n), i.e. $T_n^{(a)} = a^{-1} T_n$, and (ii) the point process (T_n) satisfies some mixing condition which, for example, is fulfilled for Markov modulated Poisson processes and for recurrent point processes (cf. [11], [12]).

Throughout this section we use the variable γ in the following sense:

$$\gamma = \frac{\alpha}{2} \quad \text{for PS} \quad \text{and} \quad \gamma = \frac{\alpha}{4} \quad \text{for GS.}$$

4.1 Light-Traffic Derivatives via Little's and Brumelle's Formulas

Because Little's and Brumelle's formulas can be seen as special cases of Campbell's formula for stationary point processes (cf. [16]), the present section is closely related to the light-traffic approaches considered in [3] and [5]. First we investigate the limit behavior of the stationary mean waiting time $\boldsymbol{E}_0 V$ of a random customer when the arrival intensity a tends to zero.

Theorem 3 We have

$$\lim_{a \to 0} \boldsymbol{E}_0 V = \gamma. \tag{4.3}$$

Proof Clearly,

$$\boldsymbol{E}_0 V = \boldsymbol{P}_0(A_\alpha)\, \boldsymbol{E}_0(V | A_\alpha) + \boldsymbol{E}_0(V I_{A_\alpha^c}), \tag{4.4}$$

where

$$\boldsymbol{E}_0(V | A_\alpha) = \gamma, \tag{4.5}$$

for all a satisfying (2.3). Furthermore, from (4.1) and (4.2) we get that

$$\lim_{a\to 0} \boldsymbol{E}_0(VI_{A_\alpha^c}) = 0, \tag{4.6}$$

so that (4.3) follows from (4.4)–(4.6). □

The following result is an immediate consequence of Theorem 3 and of Little's and Brumelle's formulas (2.15) and (2.16).

Corollary 3 We have the light-traffic derivatives

$$\lim_{a\to 0} \frac{\boldsymbol{E}\,|W|}{a} = e_1 + \gamma \tag{4.7}$$

and

$$\lim_{a\to 0} \frac{\boldsymbol{E}\,Z}{a} = \gamma e_1 + \frac{e_2}{2}. \tag{4.8}$$

Thus, comparing the formulas in (4.7) and (4.8) with (2.13) and (2.14), we see that for the greedy server the considered light-traffic derivatives are significantly smaller than those for the polling server, in particular when service times are not too large.

Furthermore, we conjecture that, for a large class of arrival processes, the performance characteristics $\boldsymbol{E}\,|W|$ and $\boldsymbol{E}\,Z$ seen as functions of the arrival intensity a are continuously differentiable in the interval $(0, (e_1+\frac{\alpha}{2})^{-1})$. Note that, for the polling server with Poisson arrival process, this smoothness property of $\boldsymbol{E}\,|W|$ and $\boldsymbol{E}\,Z$ follows directly from (2.13) and (2.14). Then, with respect to $\boldsymbol{E}\,|W|$ and $\boldsymbol{E}\,Z$, the greedy server would be better than the polling server in a certain interval $(0, \delta)$ of positive length $\delta > 0$. However, it seems to be difficult to determine δ analytically although, clearly, it would be extremely interesting to clarify how δ depends on the form of the distributions of the sequences of arrival epochs and service times. A further open problem seems to be how (say, in case of Poisson arrivals) higher-order moments of the distribution of service times affect the performance characteristics $\boldsymbol{E}\,|W|$ and $\boldsymbol{E}\,Z$ of the greedy server. Perhaps, a simulation study could help to solve this question.

Finally, let us discuss the light-traffic behavior of the mean work load $\boldsymbol{E}\,\hat{Z}$ defined in Section 2.2. For investigating $\boldsymbol{E}\,\hat{Z}$, we introduce the following auxiliary "usual" single-server queue by considering "ficticious" service times (and omitting, in this way, vacations of the server). Namely, we add to the usual service time of an arriving customer the amount of walk time which additionally arises in the serving the actual configuration of customers on the circle (now including the newly arriving one). Of course, these ficticious

service times are not independent of the arrival process. But Brumelle's formula can still be applied. Namely, assume that the auxiliary queue is in steady state. By J, let us denote the additional walk time caused by a random arriving customer. Its service time we denote by S. Then, Brumelle's formula gives (see, e.g., Theorem 4.2.1 in [16])

$$\boldsymbol{E}\,\hat{Z} = a\,\boldsymbol{E}_0\left[V(S+J) + \frac{1}{2}(S+J)^2\right]. \tag{4.9}$$

Because $\lim_{a\to 0} \boldsymbol{E}_0[V(S+J)] = 0$ and

$$\lim_{a\to 0} \boldsymbol{E}\,(S+J)^2 = \begin{cases} e_2 + 2e_1\frac{\alpha}{2} + \frac{\alpha}{3} & \text{for PS} \\ e_2 + 2e_1\frac{\alpha}{4} + \frac{\alpha}{12} & \text{for GS.} \end{cases}$$

we get from (4.9) that

$$\lim_{a\to 0} \frac{\boldsymbol{E}\,\hat{Z}}{a} = \begin{cases} \frac{e_2}{2} + \frac{\alpha}{12}(6e_1+2) & \text{for PS} \\ \frac{e_2}{2} + \frac{\alpha}{12}(3e_1+\frac{1}{2}) & \text{for GS.} \end{cases} \tag{4.10}$$

Thus, for the greedy server, the light-traffic derivative of $\boldsymbol{E}\,\hat{Z}$ can also be significantly smaller than that for the polling server.

4.2 Light-Traffic Derivatives for Distribution of Work Load

In addition to the light-traffic derivatives (4.8) and (4.10) for the stationary mean workloads $\boldsymbol{E}\,Z$ and $\boldsymbol{E}\,\hat{Z}$ now we give corresponding light-traffic derivatives for the *distributions* of Z and $\hat{Z}$, respectively. For queueing systems without server vacations, a similar result has been obtained in [32].

First we consider the conditional distribution of the work load Z under the condition that the server is busy with serving a customer, i.e. that $W(\{0,1\}) > 0$. From the rate conservation law for stationary processes with stationary embedded point process, it follows that (see, e.g., formula (2.1) in [26])

$$P(Z > x | W(\{0,1\}) > 0) = 1 - \int_0^x P_0(Z \le x-u)\, dF_e(u), \tag{4.11}$$

for every $x \le 0$, where P_0 denotes the Palm probability measure taken with respect to the stationary point process of arrival epochs, and F_e is the distribution function of stationary residual service time given by

$$F_e(x) = e_1^{-1}\int_0^x (1-F(u))\, du. \tag{4.12}$$

From (4.1) we get

$$\lim_{a\to 0} P_0(Z \le x) = 1, \tag{4.13}$$

for every $x \ge 0$. Thus,

$$\lim_{a\to 0} P(Z > x | W(\{0,1\}) > 0) = 1 - F_e(x). \tag{4.14}$$

Furthermore, from (4.11) and (4.12) we get

$$\boldsymbol{E}\,(Z|\text{ server is busy}) = \boldsymbol{E}_0 Z + \frac{e_2}{2e_1}.$$

Assume now, additionally to (4.2), that

$$Z \le \tilde{V} \text{ and } \hat{Z} \le \tilde{V} \quad \text{for all sufficiently small } a \tag{4.2$'$}$$

which is fulfilled for the same class of arrival processes that satisfy (4.2). Then $\lim_{a\to 0} \boldsymbol{E}_0 Z = 0$ and, consequently,

$$\lim_{a\to 0} \boldsymbol{E}\,(Z|\text{ server is busy}) = \frac{e_2}{2e_1}. \tag{4.15}$$

which is in accordance with the intuition. Moreover, because

$$P(\text{ server is busy }) = \rho$$

for all sufficiently small a, from (4.8) and (4.15) we get that

$$\lim_{a\to 0} \frac{\boldsymbol{E}\,(Z|\text{ server is idle })}{(a\gamma)e_1} = 1. \tag{4.16}$$

Now, let us discuss the work load $\hat{Z}$. For this purpose, we again consider the auxiliary single-server queue with "fictive" service time $J + S$ of a random customer introduced in Section 4.1. Then, Takacs' formula on the relationship between virtual and actual workload in stationary single-server queues (see e.g. Theorem 4.5.1 of [16]) gives

$$\begin{aligned} &P(\hat{Z} > x) \\ &= a\boldsymbol{E}_0(J+S)\left(1 - \frac{1}{\boldsymbol{E}_0(J+S)}\boldsymbol{E}_0 \min\left(J+S, (x-\hat{Z})_+\right)\right) \end{aligned} \tag{4.17}$$

for every $x \ge 0$, where $y_+ = \max(0, y)$. From (4.1) and (4.2') it follows that

$$\begin{aligned} \lim_{a\to 0} \boldsymbol{E}_0 \min\left(J+S, (x-\hat{Z})_+\right) &= \lim_{a\to 0} \boldsymbol{E}_0 \min(J+S, x) \\ = \lim_{a\to 0} \int_0^x P_0(J+S > u)\,du &= \int_0^x P_0(J_0+S > u)\,du, \end{aligned}$$

where the random variable J_0 is independent of S, and uniformly distributed on $[0, \alpha]$ for the polling-server model, and on $[0, \frac{\alpha}{2}]$ for the greedy-server model, respectively. Moreover, $\lim_{a\to 0} \boldsymbol{E}_0(J+S) = \boldsymbol{E}_0(J_0+S) = \boldsymbol{E}_0 J_0 + e_1$. Thus, we get the light-traffic derivative

$$\begin{aligned} &\lim_{a\to 0} \frac{P(\hat{Z} > x)}{a} \\ &= (\boldsymbol{E}_0 J_0 + e_1)\left(1 - \frac{1}{\boldsymbol{E}_0 J_0 + e_1} \int_0^x P_0(J_0 + S > u)\, du\right), \end{aligned} \tag{4.18}$$

where $\boldsymbol{E}_0 J_0 = \gamma$. This is in accordance with the light-traffic derivative (4.10) for the expectation $\boldsymbol{E}\hat{Z}$.

References

[1] Altman, E. and Yechiali, U., Cyclic Bernoulli polling. Zeitschrift f. Operations Res. 37(1993)

[2] Altman, E. and Levy, H., Queueing in space. Preprint. INRIA Centre Sophia Antipolis/Tel-Aviv University, 1993.

[3] Baccelli, F. and Brémaud, P., Virtual customers in sensitivity and light traffic analysis via Campbell's formula for point processes. Adv. Appl. Prob. 25(1993), 221-234.

[4] Bertsimas, D.J., van Ryzin, G., A stochastic and dynamic vehicle routing problem in the Euclidean plane. Operations Res. 39(1991), 601-615.

[5] Blaszczyszyn, B., Factorial moment expansions for stochastic systems. Preprint. Wroclaw University, 1993.

[6] Borovkov, A. and Schassberger, R., Ergodicity of a polling network. Preprint. University of Novosibirsk/Technical University of Braunschweig, 1992.

[7] Boxma, O.J. and Weststrate, J.A., Waiting times in polling systems with Markovian server routing. In: Stiege, G. and Lie, J.S. (eds.), Messung, Modellierung und Bewertung von Rechnersystemen und Netzen. Springer, Berlin 1989, 89-104.

[8] Coffman, Jr. E.G. and Gilbert, E.N., A continuous polling system with constant service times. IEEE Trans. Inf. Th. IT-32(1986), 584-591.

[9] Coffman, Jr. E.G. and Gilbert, E.N., Polling and greedy servers on a line. QUESTA 2(1987), 115-145.

[10] Coffman, Jr., E.G. and Stolyar, A., Continuous polling on graphs. Probab. Engin. Inf. Sci. 7(1993), 209-226.

[11] Daley, D. and Rolski, T., Finiteness of waiting-time moments in general stationary single-server queues. Ann. Appl. Prob. 2(1992), 987-1008.

[12] Daley, D. and Rolski, T., Light traffic approximations in general stationary single-server queues. Preprint. Australian National University/Unitversity of Wroclaw, 1992.

[13] Doshi, B.T., Conditional and unconditional distributions for M/G/1 type queues with server vacations. QUESTA 7(1990), 229-252.

[14] Doshi, B.T., Generalization of the stochastic decomposition for single-server queues with vacations. Stochastic Models 6(1990), 307-333.

[15] Fendick, K.W. and Whitt, W., Measurments and approximations to describe the offered traffic and predict the average workload in a single-server queue. Proc. IEEE 77(1989), 171-194.

[16] Franken, P., König, D., Arndt, U. and Schmidt, V., Queues and Point Processes. J. Wiley & Sons, Chichester 1982.

[17] Fuhrmann, S.W. and Cooper, R.B., Stochastic decomposition in the M/G/1 queue with generalized vacations. Operations Res. 33(1985), 1117-1129.

[18] Fuhrmann, S.W. and Cooper, R.B., 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(1985), 1091-1098.

[19] Keilson, J. and Servi, L.D., Oscillating random walk models for GI/G/1 vacation systems with Bernoulli schedules. J. Appl. Prob. 23(1986), 790-802.

[20] Kleinrock, L. and Levy, H., The analysis of random polling systems. Operations Res. 36(1988), 716-732.

[21] König, D. and Schmidt, V., Random Point Processes. Teubner, Stuttgart 1992 (Germ.)

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

[23] Kroese, D.P. and Schmidt, V., Queueing systems on a circle. Zeitschrift f. Operations Res. 37(1993), 303-331.

[24] Kroese, D.P. and Schmidt, V., Single-server queues with spatially distributed arrivals. Preprint. University of Twente/University of Ulm, 1993.

[25] Lucantoni, D.M., Meier-Hellstern, K.S. and Neuts, M.F., A single-server queue with server vacations and a class of non-renewal arrival processes. Adv. Appl. Prob. 22(1990), 676-705.

[26] Miyazawa, M., Further results on decomposition formulas for single server queues with general server vacations. Preprint. Science University of Tokyo, 1992.

[27] Nahimas, S. and Rothkopf, M.H., Stochastic models of internal mail delivery systems. Management Sci. 30(1984), 1113-1120.

[28] Reiman, M.I. and Simon, B., An interpolation approximation for queueing systems with Poisson input. Operations Res. 36(1988), 454-469.

[29] Reiman, M.I. and Simon, B., Light traffic limits of sojourn time distributions in Markovian queueing networks. Stoch. Models 4(1988), 191-233.

[30] Reiman, M.I. and Simon, B., Open queueing systems in light traffic. Math. Oper. Res. 14(1989), 26-59.

[31] Schassberger, R., Stability of polling networks with state-dependent server routing. Preprint. TU Braunschweig, 1993.

[32] Sigman, K., Light traffic for workload in queues. QUESTA 11(1992), 429-442.

[33] Simon, B., Calculating light traffic limits for sojourn times in open Markovian queueing systems. Stochastic Models 9(1993), 213-231.

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

[35] Whitt, W., A light-traffic approximation for single-class departure processes from multi-class queues. Management Sci. 34(1988), 1333-1346.

[36] Whitt, W., An interpolation approximation for the mean workload in a GI/G/1 queue. Operations Res. 37(1989), 936-952.