MATHEMATICS OF OPERATIONS RESEARCH
Vol. 21, No. 1, February 1996
Printed in U.S.A.

LIGHT-TRAFFIC ANALYSIS FOR QUEUES WITH SPATIALLY DISTRIBUTED ARRIVALS

DIRK P. KROESE AND VOLKER SCHMIDT

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 efficiently than the cyclically polling server.

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. Altman and Levy (1993), Bertsimas and van Ryzin (1991), Coffman and Gilbert (1986, 1987), Coffman and Stolyar (1993), Fuhrmann and Cooper (1985b), Kroese and Schmidt (1992, 1993, 1994), Nahimas and Rothkopf (1984). 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., Takagi (1986)).

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. Coffman and Gilbert (1986), Coffman and Stolyar (1993), Fuhrmann and Cooper (1985b), Kroese and Schmidt (1992)). 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 Altman and Yechiali (1993), Blanc and van der Mei (1995), Keilson and Servi (1986), Kleinrock and Levy (1988), for Bernoulli polling, and Borovkov and Schassberger (1994), Boxma and Weststrate (1989), Schassberger (1993) for more general Markov polling. The *Brownian server* model, considered in Kroese and Schmidt (1994), 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

Received July 20, 1993; revised June 25, 1994.
AMS 1991 subject classification. Primary 60K25; Secondary: 90B22.
OR/MS Index 1978 subject classification. Primary: 681 Queues.
Key words. Queueing theory, continuous polling system, greedy server, Poisson arrivals, light-traffic derivative, Taylor-expansion, queue length, workload, general stationary input.

0364-765X/96/2101/0135/$01.25

Markov polling systems with many stations. 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, one given that the server is walking (i.e., not serving) and the other given that all customers arrive at one and the same point (i.e., no walking times). 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 the first summand can be determined, a performance analysis of the continuous polling system as a whole is possible (see, e.g., Fuhrmann and Cooper (1985b), Kroese and Schmidt (1994)). Note that analogous decomposition results hold not only for mean values, but for the corresponding distributions as well (see Boxma and Weststrate (1989), Doshi (1990a, b), Fuhrmann and Cooper (1985a), Miyazawa (1994)).

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. Kroese and Schmidt (1994)).

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 Coffman and Gilbert (1987), 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 §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 five moments $e_1, \ldots, e_5$. 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 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. From the point of view of the server, customers arrive according to a two-dimensional homogeneous Poisson process on $\mathbb{R}_+ \times [0, 1]$ with intensity a. An arbitrary atom of this process (or random measure, rather) corresponds to the time of arrival of a customer, and to his initial position on the circle with respect to the server (whom we think at 0 or 1). Once having arrived, the customers in the polling server case move towards 0 at speed α^{-1}, unless there is a customer being served, in which case all customers stop for a random (service) time. (Think of the customers "glued onto a rotating circle"). Figure 2 of Kroese and Schmidt (1992) gives an illustration. Hence, we are dealing with a kind of particle system, where particles are born at certain times and certain positions, move around and then die. We will define several performance measures of the polling and greedy server models in this more abstract and perhaps simpler context.

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. As explained before, the positions of the particles at a given time can be interpreted as the positions of the customers at this time (with respect to the server) in the polling server case.

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. This particle system describes the positions of the customers (with respect to the server) in the greedy server case.

For both particle systems 1a and 1b, let W_t denote the *random counting measure* on the Borel σ-algebra $\mathscr{B}([0,1])$ 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 Kroese and Schmidt (1992) (see also Kroese and Schmidt (1993, 1994)) 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 N with generating function

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

The link between these particle systems and our original models is somewhat more difficult than before. These systems describe the positions of the customers (with respect to the server) for both models at "traveling times" (when no service is going on). Hence the times when the server is busy are omitted. During such a (service) time one or more customers could arrive, i.e., the number of customers arriving during a service is a random variable, with generating function given by (2.1). For an illustration, see Figure 3 of Kroese and Schmidt (1992).

Let Q_t denote the random counting measure on $\mathscr{B}([0,1])$ 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 and 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. Notice that R_t describes the workload at *traveling time* t (t is the time according to a clock that stops whenever the server is busy). Z_t, however, is the workload at a "general" time t, which can be either a traveling time or a time where the server is busy. In the last case the customer at 0 has a remaining service time which is less than what it was before this service began. 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. Doshi (1990b), Lucantoni et al. (1990), Miyazawa (1994)). 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 $\mathscr{B}([0,1])$, abbreviate $M([0,1])$ to $|M|$, and, consequently, $\mathbf{E}M([0,1])$ to $\mathbf{E}|M|$.

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 Kroese and Schmidt (1994) 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 §4 in Kroese and Schmidt (1994)). However, under light traffic, stability of the greedy server follows from the following argument. Assume that

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

where α is the time which the server needs for walking once around the whole circle

(without stopping on the way). Then, using the point-process approach to Loynes' scheme concerning the construction of stationary ergodic queueing processes in (standard) single-server queues with a general stationary ergodic input (but without walking), one can show that for the greedy server the same stability theorem holds as for the polling server. The crucial idea of this approach is to construct the queueing process on the canonical probability space of the input that both the queueing process and the input are *jointly* stationary, where a coupling argument is used. In connection with this it suffices to show that, with probability one, the queue empties infinitely often for an arbitrary initial state. It is well known that the standard single-server queue possesses this property and, consequently, the desired construction exists provided that the mean service time is less than the mean interarrival time (see, e.g., Chapter 2 of Franken et al. (1982)). Next observe that, adding the value $\alpha/2$ to the service time of each customer, a stable (standard) $M/G/1$ queue is obtained 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. Thus, under (2.3), the greedy server empties infinitely often and Loynes' scheme for constructing the stationary ergodic queueing process $(\tilde{W}_t)$ works. The stationary ergodic process $(\tilde{Q}_t)$ can be constructed in a completely analogous way. 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, for a small enough, by the same consideration as above one can show that Loynes' scheme works for constructing the stationary ergodic process $(\tilde{W}_t)$, also under the assumption that the arrival epochs form a general stationary ergodic point process, for both the polling and greedy servers. This case will be considered in §4.

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 Franken et al. (1982)) it follows that

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

and

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

with probability one. Thus, the stationary mean measures $\mathbf{E}W(\cdot)$ and $\mathbf{E}Q(\cdot)$ 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, one can show for both the polling and greedy servers that, for a small enough, there exists a stationary ergodic 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

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

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 the 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

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

where $S_1, S_2, \ldots$ is a sequence of i.i.d. nonnegative 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*

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

and

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

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

For a rigorous *proof* of (2.8) we refer to Theorem 3.3 of Kroese and Schmidt (1992); (2.9) can be proved along similar lines. For a discussion of workload decomposition in polling models (with finitely many service stations), see, e.g., Boxma (1989), Boxma and Groenendijk (1987), Doshi (1990a, b), Miyazawa (1994). More "informal" proofs can be found in Boxma and Westrate (1989) and Fuhrmann and Cooper (1985a).

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 Miyazawa (1994) 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 because this increment consists not only of a service time, but for some configurations of waiting customers, also of an increase of total walking time which is not independent of the actual value of $(\hat{Z}_t)$. Consequently, the PASTA argument

used in Miyazawa (1994) does not work in our case because this argument is based on the assumption that the increment of the virtual work load process at an arrival epoch is independent of the actual value of this process.

COROLLARY 1. *In particular, it holds that*

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

and

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

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 Kroese and Schmidt (1992)). 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 $\mathbf{E}Q(\cdot)$ are usually much easier to find. The performance characteristics $\mathbf{E}|W|$ and $\mathbf{E}Z$ then follow from (2.10) and (2.11).

REMARK 1. For the polling server, the mean measure of Q is given by (see Kroese and Schmidt (1992) or Fuhrmann and Cooper (1985b)):

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

and hence,

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

and

$$\mathbf{E}Z = \frac{ae_2 + a\alpha e_1}{2(1-\rho)}. \tag{2.14}$$

Another performance characteristic, the stationary mean sojourn time of a "random customer" $\mathbf{E}_0 U$, follows from Little's formula (see, e.g., Theorem 4.2.1 of Franken et al. (1982)):

$$\mathbf{E}|W| = a\mathbf{E}_0 U. \tag{2.15}$$

Here the expectation $\mathbf{E}_0$ 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 Franken et al. (1982)):

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

where $\mathbf{E}_0 V = \mathbf{E}_0 U - e_1$ 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 §4.2 of Franken et al. (1982)). This will be used in §4 for getting light-traffic derivatives for $\mathbf{E}|W|$, $\mathbf{E}Z$ and $\mathbf{E}\hat{Z}$ 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 $\mathbf{E}|Q|$ in the Poisson arrival case. This makes it impossible to give, in this case, closed formulas for the performance characteristics $\mathbf{E}|W|$, $\mathbf{E}_0V$ and $\mathbf{E}Z$. 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 §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 $\mathbf{E}|W|$ and $\mathbf{E}Z$. 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. Kroese and Schmidt (1993)). 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. 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. We can see this in the following theorem, where a (second-order) Taylor-expansion (at $a = 0$) for $\mathbf{E}Q([0, x])$ is given, for all $x \in [0, 1]$, viewing this quantity as a function of a.

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

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

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 $\mathbf{E}Q(\cdot)$ 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.

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*:

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

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

$$\mathbf{E}Z = a\left(\frac{e_2}{2} + \frac{e_1\alpha}{4}\right) + a^2\left(\frac{e_1e_2}{2} + \frac{\alpha e_1^2}{4} + \frac{\alpha^2 e_1}{48}\right) + O(a^3). \tag{3.4}$$

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. Also if *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 Fendick and Whitt (1989), Reiman and Simon (1988a), Whitt (1989). 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 Baccelli and Brémaud (1993), Blaszczyszyn (1995) and Reiman and Simon (1989) (see also Baccelli and Brémaud (1994), Reiman and Simon (1988b), Sigman (1992), Simon (1993), Whitt (1988, 1989)). In particular, we will use an extension of Corollary 5.2 in Blaszczyszyn (1995) to the case of an independently marked Poisson process. Since we are now concerned with stationary characteristics, we assume, in view of §2, that the arrival times $\{T_n\}$ form a Poisson process on the whole real line $\mathbb{R}$ with intensity a, where we make the convention

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

Assume also that (2.3) holds. Observe that we are in the same framework as Reiman and Simon (1989), that is, we are dealing with a marked Poisson process $\{(T_n, Z_n)\}$ on $\mathbb{R}$, 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 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 (nonmarked) point process. For such processes, let W_t be the random measure of waiting customers at time t, $t \in \mathbb{R}$. And, for $t \in \mathbb{R}$, let

$$L_t = W_t([z, z + dz])\, I_{\{W_t(([0,1)) = 0\}}, \tag{3.5}$$

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 $\mathbf{E}L_0$, 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 $\mathbb{R}$, at time s. Define $G(s,x) = L_x$ and let $g(s,x) = \mathbf{E}G(s,x)$. 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 $h(s,t,x) = \mathbf{E}H(s,t,x)$. We will show by using the following Theorem 2 (which is an extension of Corollary 5.2 in Blaszczyszyn (1995)) 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 (2.3) holds and that the fifth moment e_5 of service times is finite. Moreover, Corollary 5.2 in Blaszczyszyn (1995) provides a way to calculate these derivatives via the functions h and g, defined above (see also Theorem 2 of Reiman and Simon (1989), where an additional "admissibility" condition is assumed). Related results for the "usual" multi-server queue (without walking times) have recently been derived in Blaszczyszyn, Frey and Schmidt (1995); see also the survey given in Blaszczyszyn, Rolski and Schmidt (1995). 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 $K = [0,1] \times \mathbb{R}_+$. 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 $\mathbb{R} \times K$ 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 $t \in \mathbb{R}$, 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 $t \in \mathbb{R}$ and $z \in K$, let

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

where $\delta_{(t,z)} \in \Omega_K$ denotes the counting measure with

$$\delta_{(t,z)}(B \times C) = \begin{cases} 1 & \text{for } (t,z) \in B \times C, \\ 0 & \text{for } (t,z) \notin B \times C. \end{cases}$$

Let $k \geq 1$ be an arbitrary, but fixed integer. For any $t_1, \ldots, t_k \in \mathbb{R}$ and $z_1, \ldots, z_k \in K$, let $\psi_{(t_1,z_1),\ldots,(t_k,z_k)}$ be defined by iteration of the mapping $\psi \to \psi_{(t,z)}$, i.e.,

$$\psi_{(t_1,z_1),\ldots,(t_k,z_k)}(\omega) = \big(\ldots(\psi_{(t_1,z_1)})_{(t_2,z_2)}\cdots\big)_{(t_k,z_k)}(\omega).$$

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

$$\psi_{(t_1,z_1),\ldots,(t_k,z_k)}(\omega)$$

$$= \begin{cases} \displaystyle\sum_{j=0}^{k} (-1)^{k-j} \sum_{\pi \in \left\{\binom{k}{j}\right\}} \psi\Big(\omega|_{t_k} + \sum_{i \in \pi} \delta_{(t_i,z_i)}\Big) & \text{for } t_k < \cdots < t_1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.6}$$

where $\left\{\binom{k}{j}\right\}$ denotes the collection of all those subsets of $\{1,\dots,k\}$ containing j elements.

Following Blaszczyszyn (1995), we call the functional ψ *continuous at infinities* if

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

and

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

for all $\omega,\nu\in\Omega_K$ with $\nu(\mathbb{R}\times K)<\infty$.

The following result is a straightforward extension of Corollary 5.2 in Blaszczyszyn (1995), where the nonmarked case has been considered. However, in order to make the paper more self-contained we sketch the proof of Theorem 2 in the Appendix.

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. Let $k\ge 1$ be a fixed integer. If the functional ψ is continuous at infinities, if*

$$\int_{(\mathbb{R}\times K)^i}\int_{\Omega_K}\left|\psi_{(t_1,z_1),\dots,(t_i,z_i)}(\omega)\right|P_a(d\omega)\,dt_1\,\tilde F(dz_1)\dots dt_i\,\tilde F(dz_i)<\infty \tag{3.7}$$

for all $i=1,\dots,k$, and if

$$\limsup_{a\to 0}\int_{(\mathbb{R}\times K)^{k+1}}\int_{\Omega_K}\left|\psi_{(t_1,z_1),\dots,(t_{k+1},z_{k+1})}(\omega)\right|P_a(d\omega)\cdot dt_1\,\tilde F(dz_1)\dots dt_{k+1}\,\tilde F(dz_{k+1})<\infty, \tag{3.8}$$

then

$$\mathbf{E}\psi(\{(T_n,Z_n)\})=\psi(\mathbf{0})+\sum_{i=1}^{k}a^i\int_{(\mathbb{R}\times K)^i}\psi_{(t_1,z_1),\dots,(t_i,z_i)}(\mathbf{0})\cdot dt_1\,\tilde F(dz_1)\dots dt_i\,\tilde F(dz_i)+O(a^{k+1}). \tag{3.9}$$

PROOF OF THEOREM 1. We apply Theorem 2 to the functional $\psi=L_0$ defined in (3.5). It is easy to see that this functional is continuous at infinities. Furthermore, from (3.5) and (3.6) we get that condition (3.7) is fulfilled for $k=2$ because the fourth moment e_4 of service times is finite. This can be seen in the following way: With the notation $|L|=|W_{t_i}|I_{\{W_{t_i}(\{0,1\})=0\}}$ we first show that (3.7) holds for $i=1$. Note that, for $t_1>0$, the integrand in (3.7) vanishes and, for $t_1\le 0$, we have

$$\left|L_0(\omega|_{t_1}+\delta_{(t_1,z_1)})-L_0(\omega|_{t_1})\right|\le\sum_{l=0}^{\infty}(l+1)I_{\{|L|=l\}}I_{\{-t_1\le\Sigma_{n=1}^{l+1}S_n^*\}}$$

where $S_n^*=S_n+\alpha/2$, and $S_1,S_2,\dots$ denote the service times of the customers

served after time t_1. This gives

$$\int_{\mathbb{R}\times K}\int_{\Omega_K} \left|L_0\left(\omega|_{t_1} + \delta_{(t_1,z_1)}\right) - L_0\left(\omega|_{t_1}\right)\right| P_a(d\omega)\, dt_1\, \tilde{F}(dz_1)$$

$$\leq \sum_{l=0}^{\infty} (l+1)^2 P_a(|L| = l)\left(e_1 + \frac{\alpha}{2}\right),$$

because the distribution of $|L|$ does not depend on t_1. The sum in the last inequality is finite, because $|L|$ can be bounded by the stationary queue length in the standard $M/G/1$ queue with service times S_n^*. Thus, $\mathbf{E}|L|^2 < \infty$ in view of $e_3 < \infty$. Hence, (3.7) follows for $i = 1$. Observe that, for $i = 2$, the integrand in (3.7) again vanishes if $t_1 > 0$. Thus, assume that $t_2 < t_1 \leq 0$. Moreover, we have

$$(L_0)_{(t_1,z_1),(t_2,z_2)}(\omega) = 0$$

for each $\omega \in \{\sum_{n=1}^{|L|+2} S_n^* < -t_2\}$, because in this case

$$L_0\left(\omega|_{t_2} + \delta_{(t_2,z_2)} + \delta_{(t_1,z_1)}\right) = L_0\left(\omega|_{t_2} + \delta_{(t_1,z_1)}\right)$$

and

$$L_0\left(\omega|_{t_2} + \delta_{(t_2,z_2)}\right) = L_0\left(\omega|_{t_2}\right).$$

Consequently, using analogous bounds as for $i = 1$, we obtain the inequality

$$\int_{(\mathbb{R}\times K)^2}\int_{\Omega_K} \left|(L_0)_{(t_1,z_1),(t_2,z_2)}(\omega)\right| P_a(d\omega)\, dt_1\, \tilde{F}(dz_1)\, dt_2\, \tilde{F}(dz_2)$$

$$\leq 2\sum_{l=0}^{\infty} (l+2) P_a(|L| = l) \int_{-\infty}^{0}\int_{t_2}^{0} \mathbf{P}\left(-t_2 \leq \sum_{n=1}^{l+2} S_n^*\right) dt_1\, dt_2$$

$$= \sum_{l=0}^{\infty} (l+2) P_a(|L| = l)\mathbf{E}\left(\sum_{n=1}^{l+2} S_n^*\right)^2.$$

This sum over l is finite because $\mathbf{E}|L|^3 < \infty$ in view of $e_4 < \infty$. By similar arguments we get from $e_5 < \infty$ that (3.8) is also fulfilled for $k = 2$. Thus, the Taylor expansion (3.9) holds for $\psi = L_0$ and $k = 2$. Specifically:

$$f^{(1)}(0) = \int_{-\infty}^{\infty} g(s,0)\, ds \tag{3.10}$$

and

$$f^{(2)}(0) = 2\int_{-\infty}^{\infty}\int_{-\infty}^{s} (h(s,t,0) - g(s,0) - g(t,0))\, dt\, ds. \tag{3.11}$$

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

$$\mathbf{E}Q([z, z+dz]) = f(a)/(1-\rho), \tag{3.12}$$

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

$$f^{(1)}(0) = \alpha\left(\frac{1}{2} - z\right) dz + o(dz) \tag{3.13}$$

and

$$f^{(2)}(0) = \alpha^2\left(z^2 - \frac{4z^3}{3}\right) dz + o(dz), \tag{3.14}$$

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

Instead of evaluating the integrals (3.10) and (3.11) 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 (3.10) and (3.11). 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:

$$f^{(1)}(0) = \int_0^\infty g(0,v)\, dv \tag{3.15}$$

and

$$f^{(2)}(0) = 2\int_0^\infty \int_0^\infty (h(0,u,v) - 2g(0,v))\, dv\, du. \tag{3.16}$$

From (3.15) it is easy to see that (3.13) 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,

$$\int_0^\infty h(0,u,v)\, dv = 2\int_0^\infty g(0,v)\, dv, \quad \text{for } u \geq \alpha/2. \tag{3.17}$$

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

$$k(x,u,z,dz) := \mathbf{E}^x \int_0^\infty H(0,u,v)\, dv, \tag{3.18}$$

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,

$$k(x,u,z,dz) = k(1-x,u,1-z,dz) + o(dz). \tag{3.19}$$

Hence combination of (3.13)–(3.19) shows that (3.14) is proved if we can show that

$$\begin{aligned} &2\int_0^{\alpha/2} \int_0^{1/2} [k(x,u,z,dz) + k(x,u,1-z,dz)]\, dx\, du \\ &\quad = \alpha^2\left(1 - 2z + z^2 - \frac{4z^3}{3}\right) dz + o(dz), \end{aligned} \tag{3.20}$$

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

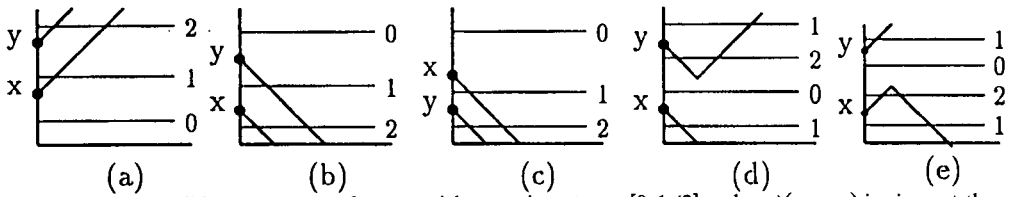


FIGURE 1. Possible movements of two particles starting at $x \in [0, 1/2]$ and y. $\lambda(x, y, z)$ is given at the end of the horizontal line-segment at height z.

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 $dz \to 0$,

$$k(x, u, z, dz) = \alpha\, dz\big(I_{[z,1/2]}(x) I_{[0,1/2]}(z) + |1/2 - z|\big) + o(dz). \tag{3.21}$$

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

$$\mathbf{E}^x \int_0^u H(0, u, v)\, dv = \alpha\, dz I_{[x-\alpha^{-1}u,\, x]}(z) + o(dz). \tag{3.22}$$

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 $x \in [0, \frac{1}{2}]$ and $y \in [0, 1]$. These particles move 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). Let

$$r(x, z) = \int_0^1 \lambda(x, y, z)\, dy. \tag{3.23}$$

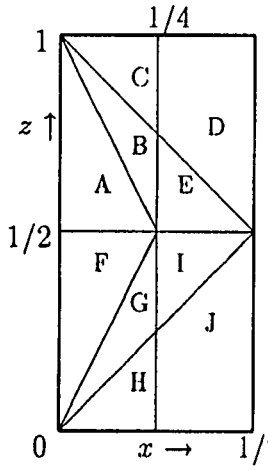
Notice that the value of $r(x, z)$ does not depend on α. A moment of reflexion will show that

$$\mathbf{E}^x \int_u^\infty H(0, u, v)\, dv = \alpha\, dz r(x - \alpha^{-1}u, z) + o(dz). \tag{3.24}$$

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.

The reader may check that from (3.22), (3.24) and Figure 2 it follows that

$$\begin{aligned} &\int_0^{1/2} \int_0^{\alpha x} k(x, u, z, dz)\, du\, dx \\ &\quad = \alpha^2\, dz(19 - 12z - 24z^2 - 40z^3)/96 + o(dz) \quad \text{if } z \in [0, 1/2], \\ &\quad = \alpha^2\, dz(-7 + 36z - 48z^2 + 24z^3)/96 + o(dz) \quad \text{if } z \in [1/2, 1]. \end{aligned} \tag{3.25}$$



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 $r(x, z)$.

On the other hand, from (3.21) it follows that for all $z \in [0, 1]$,

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

(3.26)

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

The combination of (3.19), (3.25) and (3.26) now yields the desired result (3.20), 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 $\mathbf{E}Q(\cdot)$. It is also possible to derive the expansions for $\mathbf{E}|W|$ and $\mathbf{E}Z$ directly; indeed the analysis is somewhat simpler. The Taylor-expansion for $\mathbf{E}W(\cdot)$ can also be computed via the same method.

REMARK 4. The nonmarked version of Theorem 2 has been obtained in Blaszczyszyn (1995) as a corollary of a much more general expansion for functionals of arbitrary stationary (nonmarked) 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 promising for deriving (higher-order) light-traffic approximations for queues with a (non-Poissonian) arrival process possessing some special dependence structure which still allows us to calculate limits of higher-order Palm distributions and factorial moment measures when the arrival intensity a tends to zero. In Baccelli and Brémaud (1993), conditions different from those in Blaszczyszyn (1995) and Reiman and Simon (1989) have been given for the validity of analogous formulas for light-traffic derivatives, where in Baccelli and Brémaud (1993) mostly the case has been considered that the input is a general stationary marked point process. The crucial step of the approach given in Baccelli and Brémaud (1993) is to show that a certain functional g is nonnegative almost surely (see condition (i) of Theorem 1′ in Baccelli and Brémaud (1993)) 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 Baccelli and Brémaud (1993) because their nonnegativity condition (with respect to the functionals g corresponding to these queueing characteris-

tics) seems not to hold for the greedy-server model. However, we show in §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 work-load 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 §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 small enough, 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 $\mathbf{P}_0$ of arrival epochs, finds no further customers in the system and that the next inter-arrival time is greater than x. Moreover, let V denote the stationary actual waiting time of a random customer; see also §2.4.

ASSUMPTION. We assume that

$$\lim_{a \to 0} \mathbf{P}_0(A_a) = 1 \tag{4.1}$$

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

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

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., $T_n^{(a)} = a^{-1} T_n$, which means that the mapping $a \to V$ is stochastically monotone, and (ii) the point process (T_n) satisfies some mixing condition which ensures that $\mathbf{E}_0 V < \infty$ and which, for example, is fulfilled for Markov modulated Poisson processes and for recurrent point processes (cf. Daley and Rolski (1992, 1994)).

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. Franken et al. (1982)), the present section is closely related to the light-traffic approaches considered in Baccelli and Brémaud (1993) and Blaszczyszyn (1995). First we investigate the limit behavior of the stationary mean waiting time $\mathbf{E}_0 V$ of a random customer when the arrival intensity a tends to zero.

THEOREM 3. *We have*

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

PROOF. Clearly,

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

where

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

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

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

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*

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

and

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

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 $\mathbf{E}|W|$ and $\mathbf{E}Z$ seen as functions of the arrival intensity a are continuously differentiable in the interval $(0, (e_1 + \alpha/2)^{-1})$. Note that, for the polling server with Poisson arrival process, this smoothness property of $\mathbf{E}|W|$ and $\mathbf{E}Z$ follows directly from (2.13) and (2.14). Then, with respect to $\mathbf{E}|W|$ and $\mathbf{E}Z$, 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 $\mathbf{E}|W|$ and $\mathbf{E}Z$ 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 $\mathbf{E}\hat{Z}$ defined in §2.2. For investigating $\mathbf{E}\hat{Z}$, we introduce the following auxiliary "usual" single-server queue by considering "fictitious" 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 for serving the actual configuration of customers on the circle (now including the newly arriving one). Of course, these

fictitious 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 Franken et al. (1982))

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

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

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

we get from (4.9) that

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

Thus, for the greedy server, the light-traffic derivative of $\mathbf{E}\hat{Z}$ 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 $\mathbf{E}Z$ and $\mathbf{E}\hat{Z}$ 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 Sigman (1992).

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 Miyazawa (1994)),

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

for every $x \ge 0$, 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 $x \geq 0$. 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

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

Assume now, additionally to (4.2), that

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

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

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

which is in accordance with the intuition. Moreover, because

$$P(\text{server is busy}) = \rho = ae_1$$

for all sufficiently small a, from (4.8) and from the law of total probability, taking (4.15) into account, we get that

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

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 §4.1. Then, Takacs's formula on the relationship between virtual and actual workload in stationary single-server queues (see, e.g., Theorem 4.5.1 of Franken et al. (1982)) gives

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

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

$$\lim_{a \to 0} \mathbf{E}_0 \min\left(J + S, (x - \hat{Z})_+\right) = \lim_{a \to 0} \mathbf{E}_0 \min(J + S, x)$$

$$= \lim_{a \to 0} \int_0^x P_0(J + S > u)\, du = \int_0^x P_0(J_0 + S > u)\, du,$$

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

the light-traffic derivative

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

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

Appendix. Now we sketch the proof of Theorem 2 which is a straightforward extension of Corollary 5.2 in Blaszczyszyn (1995) to the case of (independently) marked Poisson point processes. Note that, by similar arguments, Theorem 3.2 in Blaszczyszyn (1995) which concerns the factorial moment expansion of the expectation of functionals of an arbitrary nonmarked point process, can also be extended to the case of an underlying *marked* point process. However, for the purposes of the present paper, we will concentrate on functionals of a marked Poisson process.

The proof of Theorem 2 is provided by induction with respect to k and based on the following result which concerns the case $k = 0$ and which was obtained, for functionals of a general (not necessarily Poisson) point process, in Baccelli and Brémaud (1993) (see also Baccelli and Brémaud (1994) and Lemma 3.3 in Blaszczyszyn (1995)).

LEMMA A. *For a continuous at infinities functional ψ of an independently marked homogeneous Poisson process $\{(T_n, Z_n)\}$ with intensity a and mark distribution $\tilde{F}$, it holds that*

$$\mathbf{E}\psi(\{(T_n, Z_n)\}) = \psi(\mathbf{0}) + a\int_{\mathbb{R}\times K}\int_{\Omega_K}\psi_{(t,z)}(\omega)\,P_a(d\omega)\,dt\,\tilde{F}(dz) \tag{A.1}$$

provided that

$$\int_{\mathbb{R}\times K}\int_{\Omega_K}|\psi_{(t,z)}(\omega)|\,P_a(d\omega)\,dt\,\tilde{F}(dz) < \infty. \tag{A.2}$$

Observe that the integral at the right-hand side of (A.1) can be written in the form

$$\int_{\mathbb{R}\times K}\int_{\Omega_K}\psi_{(t,z)}(\omega)\,P_a(d\omega)\,dt\,\tilde{F}(dz) = \int_{\mathbb{R}\times K}\mathbf{E}\psi_{(t,z)}(\{(T_n, Z_n)\})\,dt\,\tilde{F}(dz).$$

Next, for any fixed (t, z), we apply Lemma A to the functional $\psi_{(t,z)}$ of $\{(T_n, Z_n)\}$ getting that

$$\mathbf{E}\psi(\{(T_n, Z_n)\}) = \psi(\mathbf{0}) + a\int_{\mathbb{R}\times K}\psi_{(t,z)}(\mathbf{0})\,dt\,\tilde{F}(dz)$$

$$+ a^2\int_{(\mathbb{R}\times K)^2}\int_{\Omega_K}(\psi_{(t_1,z_1)})_{(t_2,z_2)}(\omega)\,P_a(d\omega)\,dt_1\,\tilde{F}(dz_1)\,dt_2\,\tilde{F}(dz_2)$$

provided that, besides (A.2),

$$\int_{(\mathbb{R}\times K)^2}\int_{\Omega_K}\left|\left(\psi_{(t_1,z_1)}\right)_{(t_2,z_2)}(\omega)\right|P_a(d\omega)\,dt_1\,\tilde{F}(dz_1)\,dt_2\,\tilde{F}(dz_2)<\infty.$$

Then, using the fact that

$$\psi_{(t_1,z_1),\ldots,(t_i,z_i)}(\omega)=\left(\ldots\psi_{(t_1,z_1)}\ldots\right)_{(t_i,z_i)}(\omega),$$

by induction we obtain the expansion

$$\begin{aligned}
&\mathrm{E}\psi(\{(T_n,Z_n)\})\\
&\quad=\psi(\mathbf{0})+\sum_{i=1}^{k}a^i\int_{(\mathbb{R}\times K)^i}\psi_{(t_1,z_1),\ldots,(t_i,z_i)}(\mathbf{0})\,dt_1\,\tilde{F}(dz_1)\ldots dt_i\,\tilde{F}(dz_i)\\
&\quad+a^{k+1}\int_{(\mathbb{R}\times K)^{k+1}}\int_{\Omega_K}\psi_{(t_1,z_1),\ldots,(t_{k+1},z_{k+1})}(\omega)P_a(d\omega)\\
&\quad\cdot dt_1\,\tilde{F}(dz_1)\ldots dt_{k+1}\tilde{F}(dz_{k+1})
\end{aligned}$$

provided that, for all $i=1,\ldots,k+1$,

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

From this, the statement of Theorem 2 follows.

References

Altman, E., H. Levy (1993). *Queueing in space*. Preprint 4-93, Rutgers University, New Brunswick, New Jersey.

_____, U. Yechiali (1993). Cyclic Bernoulli polling. *Z. Oper. Res.* **38** 55–76.

Baccelli, F., P. Brémaud (1993). Virtual customers in sensitivity and light traffic analysis via Campbell's formula for point processes. *Adv. Appl. Probab.* **25**, 221–234.

_____, _____ (1994). *Elements of Queueing Theory*. Springer, Berlin.

Bertsimas, D. J., G. van Ryzin (1991). A stochastic and dynamic vehicle routing problem in the Euclidean plane. *Oper. Res.* **39** 601–615.

J. P. C. Blanc, J. P. C., R. D. van der Mei (1995). Optimization of polling systems with Bernoulli schedules. *Performance Evaluation* **22** 139–158.

Blaszczyszyn, B. (1995). Factorial moment expansion for stochastic systems. *Stoch. Proc. Appl.* **56** 321–335.

_____, A. Frey, V. Schmidt (1995). Light-traffic approximations for Markov-modulated multi-server queues. *Stochastic Models* **11** 423–445.

_____, T. Rolski, V. Schmidt (1995). Light-traffic approximations in queues and other stochastic models. Dshalalow, J. H., ed., *Advances in Queueing: Theory, Methods and Open Problems*. CRC Press, Boca Raton, FL, 225–242.

Borovkov, A., R. Schassberger (1994). Ergodicity of a polling network. *Stochastic Process Appl.* **50** 253–262.

Boxma, O. J. (1989). Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems Theory Appl.* **5** 185–214.

_____, W. P. Groenendijk (1987). Pseudo-conservation laws in cyclic service systems. *J. Appl. Probab.* **24** 949–964.

_____, J. A. Weststrate (1989). Waiting times in polling systems with Markovian server routing. Stiege, G. and Lie, J. S., eds., *Messung, Modellierung und Bewertung von Rechnersystemen und Netzen*. Springer, Berlin, 89–104.

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

_____, _____ (1987). Polling and greedy servers on a line. *Queueing Systems Theory Appl.* **2** 115–145.

_____, A. Stolyar (1993). Continuous polling on graphs. *Probab. Engin. Inf. Sci.* **7** 209–226.

Daley, D., T. Rolski (1992). Finiteness of waiting-time moments in general stationary single-server queues. *Ann. Appl. Probab.* **2** 987–1008.

_____, _____ (1994). Light traffic approximations in general stationary single-server queues. *Stochastic Process Appl.* **49** 141–158.

Doshi, B. T. (1990a). Conditional and unconditional distributions for $M/G/1$ type queues with server vacations. *Queueing Systems Theory Appl.* **7** 229–252.

_____ (1990b). Generalization of the stochastic decomposition for single-server queues with vacations. *Stochastic Models* **6** 307–333.

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

Franken, P., D. König, U. Arndt, V. Schmidt (1982). *Queues and Point Processes*. J. Wiley & Sons, Chichester.

Fuhrmann, S. W., R. B. Cooper (1985a). Stochastic decomposition in the $M/G/1$ queue with generalized vacations. *Oper. Res.* **33** 1117–1129.

_____, _____ (1985b). 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.

Keilson, J., L. D. Servi (1986). Oscillating random walk models for $GI/G/1$ vacation systems with Bernoulli schedules. *J. Appl. Probab.* **23** 790–802.

Kleinrock, L., H. Levy (1988). The analysis of random polling systems. *Oper. Res.* **36** 716–732.

König, D., V. Schmidt (1992). *Random Point Processes*. Teubner, Stuttgart, Germany.

Kroese, D. P., V. Schmidt (1992). A continuous polling system with general service times. *Ann. Appl. Probab.* **2** 906–927.

_____, _____ (1993). Queueing systems on a circle. *Z. Oper. Res.* **37** 303–331.

_____, _____ (1994). Single-server queues with spatially distributed arrivals. *Queueing Systems Theory Appl.* **17** 317–345.

Lucantoni, D. M., K. S. Meier-Hellstern, M. F. Neuts (1990). A single-server queue with server vacations and a class of nonrenewal arrival processes. *Adv. Appl. Probab.* **22** 676–705.

Miyazawa, M. (1994). Decomposition formulas for single server queues with vacations: A unified approach by the rate conservation law. *Stochastic Models* **10** 389–413.

Nahimas, S., M. H. Rothkopf (1984). Stochastic models of internal mail delivery systems. *Management Sci.* **30** 1113–1120.

Reiman, M. I., B. Simon (1988a). An interpolation approximation for queueing systems with Poisson input. *Oper. Res.* **36** 454–469.

_____, _____ (1988b). Light traffic limits of sojourn time distributions in Markovian queueing networks. *Stochastic Models* **4** 191–233.

_____, _____ (1989). Open queueing systems in light traffic. *Math. Oper. Res.* **14** 26–59.

Schassberger, R. (1993). *Stability of polling networks with state-dependent server routing*. Preprint. Techn. University of Braunschweig.

Sigman, K. (1992). Light traffic for workload in queues. *Queueing Systems Theory Appl.* **11** 429–442.

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

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

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

_____ (1989). An interpolation approximation for the mean workload in a $GI/G/1$ queue. *Oper. Res.* **37** 936–952.

D. P. Kroese: Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; e-mail: d.p.kroese@math.utwente.nl

V. Schmidt: Department of Stochastics, University of Ulm, Helmholtzstrasse 18, 89069 Ulm, Germany; e-mail: schmidt@mathematik.uni-ulm.de