

## QUEUEING IN SPACE

EITAN ALTMAN,\* *INRIA*  
HANOCH LEVY,\*\* *RUTCOR*

### Abstract

We consider a problem in which a single server must serve a stream of customers whose arrivals are distributed over a finite-size convex space. Under the assumption that the server has full information on the customer location, obvious service policies are the FCFS and the greedy (serve-the-closest-customer) approaches. These algorithms are, however, either inefficient (FCFS) or 'unfair' (greedy).

We propose and study two alternative algorithms, the *gated-greedy policy* and the *gated-scan policy*, which are more 'fair' than the pure greedy method. We show that the stability conditions of the gated-greedy are  $\rho < 1$  (where  $\rho$  is the expected rate at which work arrives at the system), implying that the method is at least as efficient (in terms of system stability) as any other discipline, in particular the greedy one. For the gated-scan policy we show that for any  $\rho < 1$  one can design a *stable* gated-scan policy; however, for any fixed gated-scan policy there exists  $\rho < 1$  for which the policy is unstable. We evaluate the performance of the gated-scan policy, and present bounds for the performance of the gated-greedy policy.

These results are derived for systems in which the arrivals occur on a two-dimensional space (a square) but they are not limited to this configuration; rather they hold for more complex  $N$ -dimensional spaces, in particular for serving customers in (three-dimensional) convex space and serving customers on a line.

ERGODICITY CONDITIONS; FAIRNESS; OPTIMAL STABILITY REGION; POLLING ON THE PLANE; GATED-GREEDY REGIME; GATED-SCAN REGIME; PERFORMANCE EVALUATION; BOUNDS

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60K25

### 1. Introduction

We consider a problem in which a single server must serve customers whose arrivals are distributed over a finite-size convex space. In particular we are interested in a two-dimensional space, a square plane  $D$  of dimensions  $d \times d$ . Customers arrive according to a Poisson process to the square, and the location of the arrivals is given by some distribution function. The single server moves between the customers in order to serve them. This model can represent a town (the square)

---

Received 22 March 1993; revision received 30 November 1993.

\* Postal address: INRIA Centre Sophia Antipolis, 2004 Route des Lucioles, 06560 Valbonne, France.

\*\* Postal address: RUTCOR, Rutgers University, POB 5062, New Brunswick, NJ 08903-5062, USA.

On leave from Tel-Aviv University.

and a repairman (the server) that travels between the homes in the town to apply repairs when they are demanded. Applications for this problem on a one-dimensional space are in disk operating systems. An application in three-dimensional space is the repair of satellites by a spacecraft, although at the moment this may still be in the realm of science fiction.

We investigate strategies for the efficient service of the customers by the server. We consider both a situation in which the server has full information on the exact location of every customer upon its arrival (*global information assumption*) and a situation in which the server can tell the exact location of customers only when arriving at their local neighborhood (*local information assumption*).

The obvious 'fair' discipline is the first-come first-served (FCFS) strategy, where customers are served according to the order of their arrival. This discipline, however, seems not to be efficient, since the server may have to spend much time traveling from one part of the square to the other. Indeed, we show (Section 3) that the system is not stable for all  $\rho < 1$ . It is natural therefore to consider the *serve-the-closest-customer* (*greedy*) discipline, in which, upon completing the service of a customer, the server moves to serve the closest customer. The drawback of this discipline is that it may become very unfair.

For example, consider a situation in which two customer groups, A and B, are located far away from each other and the server is serving group A customers; then group B customers have to wait not only till all group A customers are served, but also till all the customers that arrive in the vicinity of group A during the service of this group (the 'descendants' of group A) are served.

We present and study two alternative strategies: the *gated-greedy* and the *gated-scan* policies. In both policies the server operates in 'cycles'; in each cycle the server serves a customer if and only if it is present at the system at the beginning of the cycle. In the *gated-greedy* discipline the order of service within the cycle is determined by the 'greedy' approach in which the server always picks for service the closest customer to it. This policy is suitable for situations in which the server can be assumed to have global information about customer location: in other words, the server knows the exact location of every customer present. In the *gated-scan* policy at every cycle the server conducts a fixed path tour over the plane. At every point along the tour the server picks the customers which are in the near neighborhood of that point. This policy is suitable for situations in which the server *does not possess* global information about customer location, and only local information exists.

Both of the proposed gated policies are less 'unfair' than the greedy method, since the situation described in the example given above is eliminated. Moreover, we show that the *gated-greedy* discipline is stable for any  $\rho < 1$ , implying that in terms of *system stability* it is at least as efficient as the greedy discipline. The *gated-greedy* discipline can therefore be considered as an alternative which is both fair and efficient. Furthermore, since  $\rho < 1$  is proved to be a necessary condition for the

system stability, the gated-greedy policy can be considered to be an optimal policy (in terms of stability region).

As for the gated-scan policy we show that for every  $\rho < 1$  one can construct a stable gated-scan policy; however, for every gated-scan policy there exists  $\rho < 1$  for which the policy is unstable. This policy is therefore somewhat less efficient than the gated-greedy but its advantage is in not requiring global information about customer location.

We evaluate the performance of the gated-scan policy, and present bounds for the performance of the gated-greedy policy.

The analysis in this paper focuses on deriving the results for convex plane topologies, in particular a square  $D$  of dimensions  $d \times d$ . In Section 2 we describe the model. In Section 3 we derive a necessary condition for stability of any policy and prove that in terms of efficiency the first-come first-served (FCFS) policy is not optimal since  $\rho < 1$  is not a sufficient condition for stability. In Section 4 we analyze the gated-greedy policy and in Section 5 the gated-scan policy. The generalization of the results to other topologies is discussed in Section 6. In particular the results apply to: any convex finite plane, one-dimensional convex topologies (serving customers on a line), and three-dimensional convex topologies (serving customers in space). In general the results apply to any  $N$ -dimensional convex space and to some non-convex spaces as well.

1.1. *Previous work.* Greedy disciplines for polling systems have been studied only in the context of *one-dimensional* topologies [13], [17]. Other studies regarding polling systems in *continuous space one-dimensional* topologies are [1], [4], [11], [12]. In the context of cyclic polling models with  $N$  queues, Boxma et al. [9] have proposed the *globally-gated policy*, in which at every cycle the server serves all (and only) the customers which are present at the system (at all queues) at the cycle beginning. Both the gated-scan and gated-greedy policies can be considered as generalizations of the globally-gated policy to spatial topologies. Some previous results on stability of polling systems with a finite number of queues can be found in [2], [3], [14], [15], [20], [23].

1.2. *Recent related work.* After completing the writing of this paper we came across a very interesting paper by Bertsimas and Van Ryzin [8] who consider the same model analyzed in this work but emphasize other aspects of the problem. In particular, they introduce and analyze the following policies:

1. The *FCFS policy*.
2. The *partitioning policy*, in which the square is divided into smaller squares; as in our gated-scan policy, the server has a prespecified scan route that goes through these squares. In each small square the customers are served according to the FCFS discipline.
3. The *traveling salesman policy*, in which there is a central station from which the server operates; the server waits in the central station till there are at least  $n$

customers in the system, and then goes to serve them according to the optimal path, thus solving a static traveling salesman problem. Once the  $n$  customers are served, the server goes back to the central station and then goes to serve the next  $n$  customers (once they arrive) and so on.

4. The *space filling curve policy*, proposed first by Bartholdi and Platzman [6], is based on some continuous mapping from a unit circle onto the square; service is then given according to the increasing order of the preimages of the customers.

5. Some partial results are given for the *fully greedy policy*.

Bertsimas and Van Ryzin obtain lower bounds for the expected waiting time under an arbitrary policy for both light traffic and for heavy traffic conditions. They state the necessary and sufficient conditions for stability under the different policies, and study their performance.

The contributions of our work in relation to those recent results are the following:

1. The gated greedy policy proposed in our work outperforms policies 1–3 above, in the sense that its sufficient (and necessary) condition for stability is weaker, namely  $\rho < 1$ . The necessary and sufficient conditions for the stability of policies 4–5 above are still unknown.

2. Our assumptions allow for more generality. In particular, we analyze arbitrary distributions of the location of arrivals (rather than uniform), and we allow for service times that depend on customer location. We obtain necessary conditions for stability for general stationary arrival and service processes (rather than Poisson arrivals and general independent service distribution).

3. We present rigorous proofs for sufficient and necessary conditions for stability for the FCFS, the gated-greedy and the gated-scan policy. These require handling Markov chains whose state space are the set of all possible configurations of the customers on the plane, as well as the location of the server. We use novel techniques, recently introduced by Meyn and Tweedie [19], to establish the stability conditions, and in particular, geometric ergodicity. The techniques used in this paper for proving stability conditions extend easily to the partitioning policy and to the traveling salesman policy (TSP) considered in [8].

## 2. Model description and preliminaries

Consider a square  $D$  of size  $d \times d$ . We consider a Poisson arrival process with rate  $\lambda$ ; the location of an arrival is distributed according to some distribution function  $L(\cdot)$  on  $D$ . We allow for the location distribution to have a mass at some points, which may correspond to a queueing phenomenon. A customer that arrives to location  $l \in D$  requires service time which is distributed like a random variable  $B(l)$  with  $b(l)$  and  $b^{(2)}(l)$  as first and second moments. We assume that there exist some uniform positive bounds on  $b(l)$ , i.e. there exist some  $\bar{b} \geq \underline{b} > 0$  such that  $\bar{b} \geq b(l) \geq \underline{b}$  for every  $l \in D$ .

For any  $x, y \in D$ , let  $\delta(x, y)$  denote the distance between  $x$  and  $y$ . The travel time (*walking time*) from any point  $x$  to any point  $y$  is random, distributed like some random variable  $S(x, y)$ , with  $s(x, y)$  as first moment. We assume that there exists some constant  $r$  such that  $s(x, y) \leq r\delta(x, y)$  for every  $x$  and  $y$  in  $D$ . Moreover, for  $x \neq y$  we have  $s(x, y) > 0$ .

Given a function  $f: D \rightarrow \mathbb{R}$  and a measure  $\xi$  on  $D$ , define

$$f \cdot \xi = \int_D f(l)\xi(dl).$$

In particular, we shall use the notation  $|\xi| := 1 \cdot \xi = \int_D \xi(dl)$ . Using this notation we can now define the expected service time of an arbitrary customer,  $\beta := b \cdot L$  and the system utilization,  $\rho := \lambda(b \cdot L)$ . Let  $\beta^{(2)} := b^{(2)} \cdot L$  denote the second moment of the service time of an arbitrary customer.

Unless otherwise stated, we shall assume the following:

1. interarrival times are independent of the location of the arrivals;
2. given the location of the arrivals, the interarrival times, the service times and the walking times are independent;
3. the walking time along a path  $x \rightarrow y \rightarrow z$  is distributed like the sum of independent walking times  $S(x, y)$  plus  $S(y, z)$ .

For a finite set  $\mathcal{W}$ , we denote by  $|\mathcal{W}|$  the number of elements in  $\mathcal{W}$ .

We shall establish below conditions for the system stability under several policies. By stability we mean that the total expected workload  $E[V_t]$  in the system at time  $t$  is bounded in time, i.e., there exists some finite value  $v$  such that for all  $t$ ,  $E[V_t] < v$ . We shall in fact establish stability in a stronger sense (the geometric ergodicity of some embedded Markov chains).

### 3. Necessary condition for stability and the stability of the FCFS policy

In this section we derive a necessary condition for the stability of any policy for service on the plane; we show that  $\rho < 1$  is such a necessary condition. Next we analyze the first come first served (FCFS) policy when applied on the plane and derive its stability conditions; specifically, we show that there exist utilization values  $\rho < 1$  for which the FCFS policy is unstable.

#### 3.1. Necessary conditions for the stability of arbitrary policies

*Theorem 3.1.* Let  $\pi$  be an arbitrary policy for serving customers on the square plane  $D$ . A necessary condition for the stability of  $\pi$  is  $\rho < 1$ .

*Proofs.* Let System  $\mathcal{A}$  denote the system for which the claim is to be proved. To

carry out the proof let us consider an alternative system, denoted System 0, which is pathwise identical (in terms of arrival times, arrival locations and service times) to System  $\mathcal{A}$ , but in which the walking time between every two points is zero; in other words, for every  $x, y$ ,  $S(x, y) = 0$ . Also, let us consider a policy  $\pi_0$  which is a non-idling policy (namely,  $\pi_0$  serves customers as long as there are customers in the system) applied on System 0. Note that the order by which  $\pi_0$  serves the customer does not affect the amount of unfinished work in the system; we thus may assume without loss of generality that  $\pi_0$  uses the FCFS order.

Using induction, and following the approach taken in Levy et al. [18], it follows that at every moment  $t$  the amount of work at System 0 under  $\pi_0$  is less than or equal to that of System  $\mathcal{A}$  under  $\pi$ . A necessary condition for the stability of System 0 under  $\pi_0$  is therefore also a necessary condition for the stability of System  $\mathcal{A}$  under  $\pi$ .

To conclude the proof note that in System 0 the service time of the  $i$ th arriving customer is independent of its index (simply since its location is independent of its index) and thus its service time is independent of its arrival time and of the service time of the  $(i - 1)$ th customer. System 0 is therefore exactly identical in behavior to an  $M/GI/1$  system with arrival rate  $\lambda$  and with service time distributed as that of an arbitrary arriving customer at System  $\mathcal{A}$ . The expectation of the latter is given by  $\beta = b \cdot L$ . It is well known that a necessary condition for the stability of this  $M/GI/1$  system is  $\lambda\beta < 1$  (e.g. p. 238, Chapter 4 Section 2 of the second part of [10]) and thus the proof follows.

Theorem 3.1 can be generalized to arbitrary stationary arrival and service processes. Below we establish this claim and introduce the necessary notation. Denote

- $B_n$  the service time of the  $n$ th arriving customer;
- $\tau_n$  the interarrival time between the  $(n - 1)$ th and the  $n$ th arriving customer;
- $Y_n$  the location of the  $n$ th customer to be served;
- $Z_n$  the location of the  $n$ th arriving customer;
- $S_n$  the walking time between the  $(n - 1)$ th and the  $n$ th served customers.

We shall assume that  $E[\tau_1] < \infty$  or  $E[B_1] < \infty$  and

- (1)  $(B_n, \tau_n, Z_n)$  is a jointly stationary sequence, defined on a probability space  $(\Omega, \mathcal{F})$ , and which does not depend on the policy of the server discipline.

Let  $\lambda := 1/E[\tau_1]$ .  $Y_n$ , which is also defined on  $(\Omega, \mathcal{F})$ , depends of course on the service discipline; in particular, if  $\tau \ni$  FCFS service (discussed below) is used, then  $Y_n = Z_n$ . Note that we allow the random variables  $B_n, \tau_n, Z_n, \xi_n, S_n, n = 1, 2, \dots$  to be dependent, and  $\rho$  stands for  $E[B_1]/E[\tau_1]$ .

*Theorem 3.2. Consider the general stationary case (1) with arbitrary distributions*

of interarrival times, service times, walking times and locations. Let  $\pi$  be an arbitrary policy for serving customers on the square plane  $D$ . A necessary condition for the stability of  $\pi$  is  $\rho \leq 1$ .

*Proof.* The proof is the same as for the previous theorem. The only difference is that the instability of system 0 for  $\rho > 1$  follows this time from the instability of a  $G/G/1$  queue for  $\rho > 1$ , for example [5] p. 36.

**3.2. Analysis of the FCFS policy.** The FCFS policy is implemented as follows. The server serves the customers in the order of their arrival. When the service of the  $i$ th customer is completed then one of the following occurs:

1. If at that moment the  $(i + 1)$ th customer is already present in the system then the server moves to serve him immediately.
2. Otherwise, namely if the  $(i + 1)$ th customer is not present in the system at that moment, then the server waits in its current position until the  $(i + 1)$ th customer arrives, and then moves to serve him.

We show below that there exists a utilization  $\rho < 1$  for which the FCFS is unstable, under the general arrivals and service condition (1). Let  $s_{\text{FCFS}} = E[S_1]$ .

**Theorem 3.3.** *Assume that  $(B_n, \tau_n, Z_n, S_n)$  is a jointly stationary sequence, defined on a probability space  $(\Omega, \mathcal{F})$ . If  $1 - s_{\text{FCFS}}\lambda < \rho$  then the system is unstable under the FCFS policy.*

*Proof.* Let System  $\mathcal{A}$  be the system under consideration and  $\rho_{\mathcal{A}}$  be the utilization of this system such that  $1 - s_{\text{FCFS}}\lambda < \rho_{\mathcal{A}}$ . We construct System  $\mathcal{B}$  which mimics the behavior of System  $\mathcal{A}$  but which does not have walking times; rather, the walking time incurred in System  $\mathcal{A}$  for serving the  $i$ th customer is added in System  $\mathcal{B}$  to the service time of the  $i$ th customer. It is obvious that the behavior of the two systems is identical and that at every epoch  $t$  the number (and identity) of customers in both systems is the same. Thus, System  $\mathcal{A}$  is stable if and only if System  $\mathcal{B}$  is stable.

We have in the new system:  $(B_n, Z_n, \tau_n, S_n)_{\mathcal{B}} = (B_n + S_n, Z_n, \tau_n, 0)_{\mathcal{A}}$  (note that  $Z_n = Y_n$ ) which is again stationary. Moreover, in System  $\mathcal{B}$  the walking times are zero. Thus, System  $\mathcal{B}$  is a  $G/G/1$  system in which the expected service time is given by  $\beta_{\mathcal{B}} = \beta_{\mathcal{A}} + s_{\text{FCFS}}$ . The expected amount of work arriving to this system during a time unit is  $\rho_{\mathcal{B}} = (E[B_1] + s_{\text{FCFS}})/E(\tau_1) > 1$  and thus the system is unstable (see for example [5], p. 36).

#### 4. The gated-greedy service discipline

We consider in this section the following gated-greedy service discipline, which combines the FCFS and greedy properties. At time  $T(0) = 0$ , all the customers that are present at the system are tagged. The server then moves between the tagged customers using the greedy discipline, i.e. moving always to the closest customer (if any). If there are several customers at an equal distance then the server moves to

one of them according to some decision rule (e.g. with equal probability). If several tagged customers are located in the same place, we assume that they are served according to the FCFS order. (The results remain valid for any other order of service of these customers.) If the system is empty at  $T(0)$ , then the server remains in its place till the next customer arrives; this customer is then tagged, and the server moves to serve it. Let  $T(1)$  be the first time after  $T(0)$  when all the tagged customers have been served. At this time the procedure described above is repeated with  $T(1)$  replacing  $T(0)$ . This procedure is repeated, and  $T(n)$  are defined recursively.

4.1. *The stability conditions.* Our objective in this section is to show that for any  $\rho < 1$  the system is stable under the gated-greedy discipline.

The state of the system is given by two objects: a counting measure that describes the customer location configuration, and the location of the server. We consider the embedded state process  $(Q_n, U_n)$  representing these objects at time  $T(n)$ , which forms a Markov chain. Let  $X$  denote the state space of this Markov chain, and let  $\mathcal{Q}$  denote the set of counting measures on  $D$ . We shall understand  $0 \in \mathcal{Q}$  to correspond to an empty system.

*Lemma 4.1.* *Assume that there are  $N$  customers on the square  $D$ , and the server follows the greedy discipline to serve them (not serving future arriving customers). Then the total distance  $R(N)$  traveled by the server is smaller than or equal to  $\alpha\sqrt{N}$ , where*

$$(2) \quad \alpha \leq \frac{8d}{\sqrt{\pi}}.$$

*Proof.* The inequality (2) clearly holds for  $N < 3$ . We thus assume below  $N \geq 3$ . Define  $r_k$  to be the distance traveled by the server between the  $k$ th and the  $(k+1)$ th customers served. For  $0 \leq x \leq \sqrt{2}d$  let  $A_x$  be the subset of the  $N$  customers served obeying  $r_k \geq x$  for some  $k$ . Thus a customer belongs to  $A_x$  if and only if the distance traveled by the server from that customer to the next closest one is greater or equal to  $x$ . Let  $N_x = |A_x|$ . First we show that for  $0 \leq x \leq \sqrt{2}d$ ,

$$(3) \quad N_x \leq \frac{16}{\pi} \left(\frac{d}{x}\right)^2.$$

To prove (3) we first claim that if (3) *does not hold* then there are at least two points  $a, a' \in A_x$  such that the distance between them  $\delta(a, a') < x$ . (By definition of  $A_x$ ,  $a$  and  $a'$  are not served consecutively.)

For the contradiction, assume that this claim does not hold, namely that

$$N_x > \frac{16}{\pi} \left(\frac{d}{x}\right)^2$$

and that  $\delta(a, a') \geq x$  for all  $a, a' \in A_x$ . Then each of the points in  $A_x$  can draw

around itself a circle of radius  $x/2$  in which no point of the set  $A_x$  exists and the total area of these circles is  $N_x \pi x^2/4 > 4d^2$ . If  $x < d$  then at least quarter of each circle is contained in the square; this means that the area of the square is greater than  $d^2$ , which is a contradiction. If  $\sqrt{2}d \geq x > d$  then the surface of the intersection of each circle and the square is at least  $\pi d^2/4$  and thus the surface of the square is at least  $N_x \pi d^2/4 > 4(d/x)^2 d^2 \geq 2d^2$  (since  $x \leq \sqrt{2}d$ ), which is also a contradiction. Hence if (3) does not hold then there must exist two points  $a, a' \in A_x$  such that  $\delta(a, a') < x$ .

Let us now focus on these two points  $a$  and  $a'$ ; assume without loss of generality that  $a$  was served before  $a'$ . Since  $a \in A_x$  and since the greedy service discipline is used, the closest customer to  $a$  must have been at a distance greater or equal to  $x$ , which contradicts  $\delta(a, a') < x$ . Thus, by way of contradiction, (3) is established.

Finally, from (3), since  $N \geq 3$  and since  $N_x \leq N$  it follows that

$$\begin{aligned} R(N) &\leq \int_0^{\sqrt{2}d} N_x dx \\ &\leq \int_0^{4d/\sqrt{\pi N}} N dx + \int_{4d/\sqrt{\pi N}}^{\sqrt{2}d} (16/\pi)(d/x)^2 dx \\ &= 4\sqrt{\frac{N}{\pi}}d + 16\frac{d^2}{\pi} \left[ \frac{\sqrt{\pi N}}{4d} - \frac{1}{\sqrt{2}d} \right] \\ &< \frac{8d}{\sqrt{\pi}}\sqrt{N}. \end{aligned}$$

(Note that  $N \geq 3$  and hence  $4d/\sqrt{\pi N} < \sqrt{2}d$ .)

*Remark.* Assume that the arrival locations are distributed uniformly on the square. The bound obtained in Lemma 4.1 implies that the gated-greedy policy is optimal, in the sense that there exists no other policy that can serve any  $N$  customers in a total distance less than  $O(\sqrt{N})$ . This follows from the fact that the optimal policy is known to require  $O(\sqrt{N})$  in the case that arrivals locations are distributed uniformly on the square (see e.g. [7]).

We are now ready to establish the sufficient condition for the stability of the gated-greedy schedule. The main tool is the following.

*Lemma 4.2.* Consider a strongly aperiodic Markov chain  $X_n$  on a state space  $X$ . Assume that there exists a set  $\mathcal{K} \subseteq X$  and a function  $g: X \rightarrow [1, \infty)$  such that:

- (i) there exists some  $\epsilon > 0$  such that  $E[g(X_{n+1}) - g(X_n) | X_n] \leq -\epsilon g(X_n)$  for  $X_n \in \mathcal{K}^c$ ;
- (ii)  $E[g(X_{n+1}) | X_n] < \infty$  for  $X_n \in \mathcal{K}$ ;
- (iii)  $\mathcal{K}$  is a small set.

Then (a)  $X_n$  is geometrically ergodic. Moreover, (b.1)  $E[g(X)]$  is finite in steady state, and (b.2)  $E[g(X_n)]$  converges to  $E[g(X)]$  geometrically fast.

*Proof.* (a) is proved in [21] (or in [19], the Corollary to Theorem 6.2, when restricting to  $f = g$ ). (b.1) is proved in [22] and (b.2) follows from [19], the corollary to Theorem 6.2. In [19] it is required that the set  $\mathcal{K}$  be a ‘petite’ set. This is satisfied since any small set is also a petite set.

In the above lemma, a set  $\mathcal{K}$  is said to be small if there exists some positive measure  $\phi$  on  $X$ , such that for any  $\Gamma \subset X$  with  $\phi(\Gamma) > 0$  there exists  $j$  such that

$$(4) \quad \inf_{x \in \mathcal{K}} \sum_{n=1}^j P^n(x, \Gamma) > 0,$$

where  $P(x, \Gamma)$  are the transition probabilities of the Markov chain. A Markov chain  $X_n$  is said to be strongly aperiodic if there exists a set  $\mathcal{C} \subset X$ , a probability  $\nu$  with  $\nu(\mathcal{C}) = 1$ , and  $\delta > 0$  such that  $P(X_n \text{ enters } \mathcal{C} \text{ for some } n \geq 1 \mid X_0 = x) > \delta$  for all  $x \in X$ , and

$$P(x, B) \geq \delta \nu(B), \quad x \in \mathcal{C}, B \in \mathcal{B}(X).$$

*Theorem 4.3.* Assume that  $\rho < 1$ . Then  $(Q_n, U_n)$  is geometrically ergodic; the expected workload in steady state is finite and the transient expected workload  $E[b \cdot Q_n]$  converges to the steady state geometrically fast; the gated-greedy policy is stable.

*Proof.* The proof is based on the previous lemma. We show that the conditions of the lemma are indeed satisfied. We begin by establishing (i). Recall the definition of  $\alpha$  in (2), and choose some  $\epsilon > 0$  such that  $\rho + \epsilon < 1$ . Let

$$\mathcal{K} = \{(q, u) \in X : |q| \leq Z^*\}, \quad \text{where } Z^* := \left( \frac{\rho \alpha r}{(1 - \rho - \epsilon) \underline{b}} \right)^2,$$

( $|q|$  is the number of customers). Then for  $(Q_n, U_n) \in \mathcal{K}^c$  we have

$$E[b \cdot Q_{n+1} - b \cdot Q_n \mid Q_n] \leq \rho \alpha \sqrt{|Q_n|} + (\rho - 1)b \cdot Q_n \leq -\epsilon b \cdot Q_n$$

which follows from Lemma 4.1, the definition of  $\mathcal{K}$  as well as the fact that  $b \cdot Q_n \geq \underline{b} |Q_n|$ . Indeed,  $(1 - \rho - \epsilon)b \cdot Q_n \geq (1 - \rho - \epsilon)\underline{b} |Q_n| \geq \rho \alpha \sqrt{|Q_n|}$  by the definition of  $\mathcal{K}$ . Condition (i) of Lemma 4.2 easily follows by identifying  $g(q) = 1 + b \cdot q$ .

Next we establish (ii). We have for  $Q_n \in \mathcal{K}$

$$E[b \cdot Q_{n+1} \mid Q_n] \leq \rho \alpha \sqrt{|Q_n|} + \rho \bar{b} Z^* \leq \rho \alpha \sqrt{Z^*} + \rho \bar{b} Z^* < \infty$$

which establishes (ii).

Next, we establish (iii). For any  $l \in D$  and  $\bar{Q} \subset Q$ , let  $\hat{P}(l, \bar{Q})$  be the probability that the location configuration of the customers that arrive during a service time  $B(l)$ , is in  $\bar{Q}$ .

Denote  $\psi(l, t) :=$  the probability of no arrivals during the walking time of the server from  $l$  to  $t$ .  $\psi$  can be bounded by

$$\begin{aligned}
 \psi(l, t) &= E[\exp(-\lambda S(l, t))] \geq \exp(-E[\lambda S(l, t)]) \\
 (5) \qquad &= \exp(-\lambda s(l, t)) \geq \exp(-\lambda \delta(l, t)) \\
 &\geq \exp(-\lambda r d\sqrt{2}).
 \end{aligned}$$

Denote by  $\hat{\beta}(l)$  the probability of no arrivals during the service time  $B(l)$ . We have, similarly to (5),

$$(6) \qquad \hat{\beta}(l) \geq \exp(-\lambda b(l)) \geq \exp(-\lambda \bar{b}).$$

Let

$$\phi(\Gamma) := \int_{(q,s) \in \Gamma} L(ds) \hat{P}(s, dq).$$

$\phi(\cdot)$  is thus the joint distribution of the location of an arbitrary arrival, and the distribution of all arrivals during the service time of that arrival.

We show below that  $P^2((z, u), \Gamma)$  (the two-step transition probability) is greater than 0 uniformly in  $\mathcal{K}$ , from which the condition (4) for a set to be small is seen to be satisfied, with  $j = 2$ . Below,  $(0, u)$  will correspond (with some abuse of notation) to an empty system, with the server at location  $u$ .

Let  $\Gamma$  satisfy  $\phi(\Gamma) > 0$ , and let  $(z, u) \in \mathcal{K}$ . We have

$$\begin{aligned}
 P^2((z, u), \Gamma) &\geq \int_D \int_D P((z, u)(0, dt)) P((0, t), \Gamma) \\
 &\geq \int_D P((z, u)(0, dt)) \int_{(q,s) \in \Gamma} \psi(t, s) L(ds) \hat{P}(s, dq) \\
 &\geq P(Q_1 = 0 \mid Q_0 = z, U_0 = u) \exp(-\lambda r d\sqrt{2}) \phi(\Gamma) \\
 &\geq \exp(-(|Z^*| + 1)\lambda[\bar{b} + r d\sqrt{2}]) \phi(\Gamma).
 \end{aligned}$$

The first inequality above follows from the fact that

$$P^2((z, u), \Gamma) = \int_{Q,D} P((z, u)(dq, dt)) P((q, t), \Gamma).$$

To understand the second inequality, we note that if a transition occurs from  $(0, t)$  to  $\Gamma$  it means that the first arrival that occurred after the server finished serving at  $t$ , was to a point  $s$  (distributed according to  $L$ ) and the configuration  $q$  when finishing the service at  $s$  is distributed like the configuration of all the arrivals that occur during the service time  $B(s)$  plus the walking time from  $t$  to  $s$ : moreover,  $s$  and  $q$  are such that  $(q, s) \in \Gamma$ . The inequality is then obtained by considering transitions from  $(0, t)$  into a subset of  $\Gamma$ , obtained by restriction to cases where no arrivals occurred during the walking time from  $t$  to  $s$ .

The third inequality follows from (5); the last inequality follows from the fact that for  $(z, u) \in \mathcal{K}$  we have  $z \leq Z^*$ , and from (5) and (6).

Finally, the strong aperiodicity follows by arguments similar to those above, choosing  $\mathcal{C} := \{(q, u) : q = 0\}$ , and  $\nu(\{0, B_q\}) := L(B_q)$  for any Borel set  $B_q \subset \mathcal{Q}$ .

This establishes the proof of the geometric ergodicity.

This implies by Lemma 4.2 (b) that the expected workload  $E[b \cdot Q]$  in steady state (of the Markov chain  $(Q_n, U_n)$ ) is finite, and  $E[b \cdot Q_n]$  converges to  $E[b \cdot Q]$  geometrically fast. Since the workload at an arbitrary time is upper bounded by the sum of expected workloads in the beginning of the current cycle plus in the beginning of the next cycle, this implies the stability of the gated-greedy policy.

*4.2. Expected cycle duration, workload and number of customers.* Our aim in this subsection is to obtain upper bounds for the number of customers in the system and the sojourn time at arbitrary moments. To do that, we obtain bounds for the first and second moments of the cycle duration, as well as the workload and number of customers at the beginning of cycles.

Let  $C_n$  denote the duration of the  $n$ th cycle, and let  $V_n$  denote the workload in the system (i.e. the sum of service times that will be required by the customers present in the system) at the beginning of the  $n$ th cycle. When omitting the index  $n$  we shall mean the steady state versions of the above quantities.  $E^0$  will denote expectation at the time of the beginning of a cycle, in steady state.

*First moments.* It follows from Lemma 4.1 that

$$(7) \quad E[V_{n+1} \mid Q_n, V_n] \leq \rho(V_n + r\alpha\sqrt{|Q_n|}).$$

Taking expectations, we get from (7) for the steady state:

$$E^0[V] \leq \rho(E^0[V] + r\alpha E^0\sqrt{|Q|}).$$

Since  $E^0[V] \geq \underline{b}E^0(|Q|)$ , we have

$$(8) \quad \underline{b}(E^0\sqrt{|Q|})^2 \leq \underline{b}E^0(|Q|) \leq E^0[V] \leq \frac{\rho r\alpha}{1-\rho} E^0\sqrt{|Q|}.$$

Dividing by  $E^0\sqrt{(|Q|)}$  we obtain

$$(9) \quad E^0\sqrt{(|Q|)} \leq \frac{\rho r \alpha}{\underline{b}(1-\rho)},$$

and substituting in (8) we finally get

$$(10) \quad E^0(|Q|) \leq \left( \frac{\rho r \alpha}{\underline{b}(1-\rho)} \right)^2$$

and

$$(11) \quad E^0[V] \leq \frac{(\rho r \alpha)^2}{\underline{b}(1-\rho)^2}.$$

Finally, it follows that

$$E^0[C] \leq E^0[V] + \alpha r E^0\sqrt{(|Q|)} \leq \frac{(\rho r \alpha)^2}{\underline{b}(1-\rho)^2} + \frac{\rho r^2 \alpha^2}{\underline{b}(1-\rho)} = \frac{(r \alpha)^2}{\underline{b}(1-\rho)^2}.$$

Next we obtain lower bounds for the above quantities.

Let  $X$  be the location of the arrival of an arbitrary customer ( $X$  is thus distributed according to  $L$ ). Let  $\underline{s} := \inf_{y \in D} E s(y, X)$ . As an example, if arrivals are uniformly distributed on the plane and the expected walking time is proportional to the distance  $s(x, y) = r\delta(x, y)$  then  $\underline{s} \geq rd/4$ .

We get the following:

$$E^0[V] \geq \rho[E^0[V] + \underline{s}]$$

and hence

$$(12) \quad E^0[V] \geq \frac{\rho \underline{s}}{1-\rho}$$

$$(13) \quad E^0[C] \geq E^0[V] + \underline{s} \geq \frac{\underline{s}}{1-\rho}$$

and finally

$$(14) \quad E^0(|Q|) = \lambda E^0[C] \geq \frac{\lambda \underline{s}}{1-\rho}.$$

*Second moments.* To obtain an upper bound for the second moments, we shall assume below that  $(1-\rho^2)\underline{b} > 2\rho^2 r \alpha$ . (This will be satisfied if the bound on the walking speed,  $r$ , is small enough.) The amount of work found in the beginning of the  $(n+1)$ th cycle is the amount of work that arrived during  $C_n$ . It follows from (8) that

$$E[V_{n+1}] = \rho^2 E[(C_n)^2] + \lambda E[C_n] \beta^{(2)}$$

(this follows, for instance from the Auxiliary problem in [16], p. 238). Hence we have in steady state

$$\begin{aligned}
 E^0[V^2] &\leq \rho^2 E^0[(V + r\alpha\sqrt{(|Q|)})^2] + \lambda E^0[C]\beta^{(2)} \\
 (15) \quad &= \rho^2 E^0[V^2] + 2r\alpha\rho^2 E^0[V\sqrt{(|Q|)}] + (r\alpha\rho)^2 E^0(|Q|) + \lambda E^0[C]\beta^{(2)} \\
 &\leq \rho^2 E^0[V^2] + 2r\alpha\rho^2 E^0[V|Q|] + (r\alpha\rho)^2 E^0(|Q|) + \lambda E^0[C]\beta^{(2)}.
 \end{aligned}$$

Thus

$$(16) \quad (1 - \rho^2)E^0[V^2] \leq (2\rho^2 r\alpha)E^0[V|Q|] + (r\alpha\rho)^2 E^0(|Q|) + \lambda E^0[C]\beta^{(2)}.$$

Combining this with the fact that

$$E^0[V^2] = E^0\{E^0[V(Q \cdot B) | Q]\} \geq E^0(|Q| V) \underline{b}$$

we get

$$\begin{aligned}
 ((1 - \rho^2)\underline{b} - 2\rho^2 r\alpha)E^0(|Q| V) &\leq (r\alpha\rho)^2 E^0(|Q|) + \lambda E^0[C]\beta^{(2)} \\
 &\leq \left(\frac{(\rho r\alpha)^2}{\underline{b}(1 - \rho)}\right)^2 + \lambda\beta^{(2)} \frac{(r\alpha)^2}{\underline{b}(1 - \rho)^2} \\
 &= \frac{(r\alpha)^2[(\rho r\alpha)^2 + \lambda\beta^{(2)}\underline{b}]}{(\underline{b}(1 - \rho))^2}.
 \end{aligned}$$

Hence we obtain

$$E^0(|Q| V) \leq \frac{(r\alpha)^2[(\rho r\alpha)^2 + \lambda\beta^{(2)}\underline{b}]}{((1 - \rho^2)\underline{b} - 2\rho^2 r\alpha)[\underline{b}(1 - \rho)]^2}.$$

Substituting in (16) we get

$$(17) \quad E^0[V^2] \leq \left[\frac{r\alpha}{\underline{b}(1 - \rho)}\right]^2 \frac{(\rho r\alpha)^2 \rho^2 (\underline{b} + 2\rho^2 r\alpha) - \lambda \underline{b}^2 \beta^{(2)}}{(1 - \rho^2)\underline{b} - 2\rho^2 r\alpha}.$$

The bounds for  $E^0[C^2]$  and  $E^0(|Q|^2)$  are then obtained by substituting (9), (10), (11) and (17) into

$$\begin{aligned}
 (18) \quad E^0[C^2] &\leq E^0[V + \alpha r\sqrt{(|Q|)}]^2 \\
 &= E^0[V^2] + 2E^0[V]\alpha r E^0\sqrt{(|Q|)} + (\alpha r)^2 E^0(|Q|)
 \end{aligned}$$

and then

$$(19) \quad E^0(|Q|)^2 = \lambda^2 (E^0[C])^2 + \lambda E^0[C^2].$$

*Expected sojourn times and number of customers at an arbitrary time.* We can bound the expected sojourn time in a way similar to Boxma et al. [9]. The sojourn time of an arbitrary customer is upper bounded by the sum of the residual cycle time  $C_R$ , plus the duration of the next cycle  $C_N$ . The expected residual cycle time  $E[C_R]$  and the expected past cycle time  $E[C_P]$  are given by  $E^0[C^2]/2E^0[C]$ . Thus

$$E[T] \leq \frac{E^0[C^2]}{2E^0[C]} + E[C_N].$$

$C_N$  is given by the sum of the walking times plus the amount of work that arrived during the durations of the past cycle time and of the residual cycle time. Let  $Q_N$  be the number of customers in the beginning of the cycle that starts after the arrival of the arbitrary customer. Thus we have

$$\begin{aligned} E[C_N] &\leq \rho(E[C_P] + E[C_R]) + \alpha r E[\sqrt{|Q_N|}] \\ &\leq \rho(E[C_P] + E[C_R]) + \alpha r \sqrt{E(|Q_N|)} \\ &= \rho(E[C_P] + E[C_R]) + \alpha r \sqrt{\lambda E[C_P + C_R]} \\ &= \rho \frac{E^0[C^2]}{E^0[C]} + \alpha r \sqrt{\left(\lambda \frac{E^0[C^2]}{E^0[C]}\right)}. \end{aligned}$$

We thus get

$$E[T] \leq \left(\rho + \frac{1}{2}\right) \frac{E^0[C^2]}{E^0[C]} + \alpha r \sqrt{\left(\lambda \frac{E^0[C^2]}{E^0[C]}\right)}.$$

An upper bound for  $E[T]$  is now obtained by using (18) for an upper bound to  $E^0[C^2]$ , and (13) for a lower bound on  $E^0[C]$ . Finally, by Little's law, the expected number of customers in the system at steady state is given by  $\lambda E[T]$ .

4.3. *Other optimal gated policies.* We may consider other variants of the gated strategy where the server serves in each 'cycle' all the customers that arrived during the previous 'cycle' (the  $n$ th cycle is given by the time period between  $T(n - 1)$  and  $T(n)$ , where  $T(n)$  was defined in the beginning of this section). It is easily seen that the stability results and the results on the performance evaluation would hold for such policies provided that they satisfy a condition of the type introduced in Lemma 4.1, i.e. that the total walking distance required to serve  $N$  customers is bounded by  $\alpha \sqrt{N}$ , where  $\alpha$  is some constant. For the stability results it is sufficient in fact that the distance required for serving  $N$  customers should be of order  $o(N)$ , i.e. it should grow sublinearly in  $N$ . In particular we may consider the gated traveling salesman policy, i.e. a policy that serves customers in each cycle in the shortest possible path, and a gated version of the space filling curve policy [8].

### 5. The gated-scan scheme

The gated-greedy scheme requires *global knowledge* of the *exact location* of every customer on the plane. In many situations such global information is not available and the server can tell the exact location of a customer only when reaching its close neighborhood. In such situations alternative disciplines need to be employed. Such a policy is the *gated-scan* policy which we next study.

The gated-scan policy can be described as follows. The server follows a *fixed cyclic path* in which it scans the plane. The local neighborhood of the server along the fixed path can be described as a 'band', and the server can be thought of as traveling in the middle of the band. The path is planned in a way that the band covers the whole plane. An example of such a path and the associated band (applied

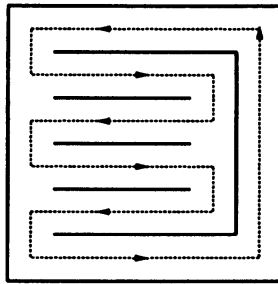


Figure 1. An example of the server's fixed path on the square

to the square) is described in Figure 1. In this figure the solid lines represent the boundaries of the band, while the dashed line represents the server's trajectory. Every point  $l$  on the plane is associated with a single point  $\sigma(l)$  on the fixed path (e.g. the point on the path which is the closest to  $l$ ), called the *corresponding fixed path point* of  $l$ .

One point on the fixed path is called the *path starting point* and is denoted  $l_0$  (note that due to the cyclic nature of the path, this is also the path end). The instant at which the server leaves the path starting point is called the *cycle beginning instant*; a *cycle* is the period between two successive cycle beginning instants. At every cycle the server serves all customers which were present on the plane at the cycle beginning instant.

The service of customers along the server's path is carried out as follows. Let  $l$  be a point on the plane in which there is a customer to be served; let  $\sigma(l)$  be the corresponding fixed path point of  $l$ . When the server arrives at  $\sigma(l)$  (along its fixed path journey), it stops its movement along the fixed path and moves to  $l$ . When arriving at  $l$ , it serves that customer, and then returns to  $\sigma(l)$ . If there are other customers requiring service in the locality of  $\sigma(l)$  (either at  $l$  itself or at other points whose corresponding fixed path point is  $l$ ), then the server will serve them in any arbitrary order; however, *for each* of these customers, the server will walk from  $\sigma(l)$  to the customer, serve the customer, and then walk back to  $\sigma(l)$ . The walk from  $\sigma(l)$  to  $l$  and back to  $\sigma(l)$  is called the *local walk*, and the time duration of such a local walk is distributed like a random variable  $\zeta(l)$ . The length of the local walk depends only on its location and on the predetermined cycle path of the server. We assume that the local (walking) times of the  $n$ th arriving customer, denoted by  $\zeta_n(Z_n)$ , are independent random variables. (Recall that  $Z_n$  is the location of the  $n$ th arriving customer.)

**5.1. The stability region.** Our objective in this section is to study the stability region of the gated-scan policy. As we will show, for any  $\rho < 1$  there exists a gated-scan policy which is stable under that  $\rho$ ; however, for any gated-scan policy there exists  $\rho < 1$  for which the policy is unstable.

First we focus on showing the first property. Given utilization  $\rho$  we will show that one can construct a stable gated-scan policy.

Let  $C_m$  be the cycle duration, and let  $Q_m$  be a counting measure that describes the customer location configuration, at the beginning of the  $m$ th cycle.

*Theorem 5.1.* For every  $\rho < 1$  there exists a stable gated-scan policy, and  $C_m$  and  $Q_m$  are geometrically ergodic Markov chains.

*Proof.* First we construct the policy. Let  $Z$  be the location of an arbitrary customer, and  $\hat{\delta} : D \rightarrow \mathbb{R}$  be given as  $\hat{\delta}(l) = \delta(l, \sigma(l))$ . The local walking time of an arbitrary customer  $\zeta$  satisfies

$$E[\zeta] \leq rE\delta(Z, \sigma(Z)) = r(\hat{\delta} \cdot L).$$

We fix an arbitrary cycle path for the server such that  $r(\hat{\delta} \cdot L) \leq (1 - \rho)\lambda^{-1}$  and hence the expected local time  $\zeta$  satisfies

$$(20) \quad E[\zeta] \leq (1 - \rho)\lambda^{-1}.$$

To prove the stability of this system, which we call System  $\mathcal{A}$ , we construct a similar system, System  $\mathcal{B}$ , in which the arrivals and service order are identical (pathwise) to those of system  $\mathcal{A}$ , but in which local times are eliminated, i.e. it takes no time for the server to travel from the fixed path point to the customer location point; this time in System  $\mathcal{B}$  will be added to the service time of the customer. In other words, in System  $\mathcal{B}$  the local travel time spent on a customer is accounted to the service of the customer (and not to traveling).

The service time at location  $l$  in System  $\mathcal{B}$  is a random variable denoted  $B_{\mathcal{B}}(l) =_d B(l) + \zeta(l)$  with mean  $b_{\mathcal{B}}(l)$ . Denote  $\rho_{\mathcal{B}} = \lambda(b_{\mathcal{B}} \cdot L)$ . Note that by our construction,  $\rho_{\mathcal{B}} < 1$ .

System  $\mathcal{A}$  and System  $\mathcal{B}$  will have completely identical behavior, so what remains to be proved is the stability of System  $\mathcal{B}$ . The duration of the  $(m + 1)$ th cycle,  $C_{m+1}$ , is distributed like the amount of work (in system  $\mathcal{B}$ ) that arrives during the  $m$ th cycle (whose duration is  $C_m$ ), plus the total scan (walking) time (in system  $\mathcal{B}$ ) during the  $(m + 1)$  cycle, which we denote by  $H_{m+1}$ . Since the fixed cycle path is identical for all cycles and since the walking along this path is independent of the number of steps made along it (see item (iii) in Section 2), it follows that  $H_m, m = 0, 1, 2, \dots$  are i.i.d.; moreover,  $H_m$  does not depend on the history prior to the  $m$ th cycle. It then follows that  $C_m$  is a Markov chain. Let  $h = E[H_m]$ . We have

$$E[C_{m+1} \mid C_m] = \rho_{\mathcal{B}}C_m + h.$$

It then follows from Lemma 4.2 that  $C_m$  is a geometrically ergodic Markov chain. (The proof is exactly the same as for the geometrical ergodicity of the standard globally-gated polling systems [9], which is given in [4], Section 6.) The claim for  $Q_m$  follows similarly by noting that

$$E[b \cdot Q_{m+1} \mid Q_m] = \rho_{\mathcal{B}}(b \cdot Q_m + h);$$

the rest then follows similarly to the proof of Theorem 4.3.

*Remark 5.2.* From the proof above we see that the transition probabilities of the Markov chain  $C_n$  are the same as the ones for standard globally-gated polling systems [9];  $\rho$  should be replaced by  $\rho_{\mathcal{B}}$  and the service time of an arbitrary customer replaced by  $B_{\mathcal{B}} \cdot L$ . It then follows from [4] Section 6, that if  $\rho_{\mathcal{B}} < 1$  then under some conditions on the Laplace–Stieltjes transform of  $H_n$  and of the service times, all moments of the cycle time exist in steady state; moreover, the transient moments converge to the steady state ones geometrically fast. The central limit theorems and the law of iterated logarithm also apply to the chain  $C_n$ .

*Theorem 5.3.* If, for some gated-scan policy  $\rho + \lambda\zeta \geq 1$ , then  $C_n$  and  $Q_n$  are not ergodic, the expected workload converges to infinity, and thus the policy is unstable.

*Proof.* Follows immediately from Theorem 3.1 applied to the system  $\mathcal{B}$  introduced in the proof of Theorem 5.1.

*5.2. Performance evaluation.* Since the transition probabilities of the Markov chain  $C_n$  are the same as the ones for standard globally-gated polling systems [9], with  $\rho$  replaced by  $\rho_{\mathcal{B}}$  and the service time of an arbitrary customer replaced by  $B_{\mathcal{B}} \cdot L$ , we can get the distribution of the cycle times by applying the results in [9] Section 2. In particular, we obtain

$$E^0[C] = \frac{h}{1 - \rho_{\mathcal{B}}}$$

where  $E^0[C]$  is the steady state expectation of a cycle, as seen by an observer that comes at the beginning of an arbitrary cycle;  $\rho_{\mathcal{B}}$  is defined in the proof of Theorem 5.1 and is given by  $E[\lambda(b \cdot L + \zeta)]$ . Denote the second moment of the service time of an arbitrary customer in system  $\mathcal{B}$  (introduced in the proof of Theorem 5.1) by  $b_{\mathcal{B}}^{(2)} := (b^{(2)} + 2b\zeta + E[\zeta^2]) \cdot L$ . We get

$$E^0[C^2] = \frac{1}{1 - \rho^2} (E[H^2] + 2h\rho_{\mathcal{B}}E^0[C] + \lambda b_{\mathcal{B}}^{(2)}E^0[C]).$$

Note that  $h$  and  $E[H^2]$  depend only on the fixed cycle path. The moments of  $\zeta$  are not always easy to compute; however, they can be bounded in the following way:  $0 \leq E[\zeta] \leq \sup_{l \in D} [r\delta(l, \sigma(l))]$ . In a similar way, one can bound the second moment of  $\zeta$ .

The moments of the number of customers at the beginning of cycles are given by

$$E^0(|Q|) = \lambda E^0[C],$$

$$E^0(|Q|^2) = \lambda^2 (E^0[C])^2 + \lambda E^0[C^2].$$

Using the first and second moments of  $C$  one can obtain the expected waiting time  $E[W^l]$  of a customer that arrives to a point  $l$  in a similar way to [9]. We shall assume below that customers who arrive in the same cycle to points  $x$  and  $t$  such that  $\sigma(x) = \sigma(t)$  will be served according to the FCFS rule.

Let the fixed cycle path of the server be represented by a closed curve. For any point  $t \in D$  which is on the cycle path, we define  $K(t)$  to be the distance along the cycle between  $t$  and the beginning of the cycle  $l_0$ . Let  $C_P$  and  $C_R$  be the past and residual cycle times; we have as in [9]

$$E[C_P] = E[C_R] = \frac{E^0[C^2]}{2E^0[C]}.$$

Then

$$\begin{aligned} E[W^l] = & E[C_R] \\ & + \lambda \int_D b_{\mathcal{B}}(t) 1\{K(\sigma(t)) < K(\sigma(l))\} L(dt) \\ (21) \quad & + \int_D s(t, t+dt) 1\{K(\sigma(t)) < K(\sigma(l))\} \\ & + \lambda L(l) b_{\mathcal{B}}(l) E[C_P]. \end{aligned}$$

As we see, the expected waiting time of an arbitrary customer that arrives to  $l$  is composed of four terms. The first one represents the expected time from the arrival of an arbitrary customer to  $l$  till the server finishes the current cycle. The rest of the terms represent the expected time it has to wait from the beginning of the next cycle till it is served. The second term represents the total expected service time of all the customers whose corresponding location on the cyclic path is nearer than that of our customer in  $l$ . The third term represents the total expected scanning time from the beginning of the cycle till the server arrives to  $\sigma(l)$ . Finally, the fourth term represents the expected service time of all the customers whose location  $x$  satisfies  $\sigma(x) = \sigma(l)$  and who arrived before our customer.

The expected sojourn time of a customer that arrives to  $l$  is obtained from (21) by the relation  $E[T^l] = E[W^l] + b(l)$ .

## 6. Serving customers on the line, and in $N$ -dimensional spaces

The results derived in Section 3 through Section 5 can be extended and generalized in several directions, which are described below.

6.1. *Arbitrary convex planes.* All the results derived so far were derived for a square  $D$  of dimensions  $d \times d$ . As we show next, all these properties also hold for any arbitrary fixed-size convex plane.

Let System  $\mathcal{A}$  be a system in which customer arrivals are over a convex plane  $P$ . To prove these properties for System  $\mathcal{A}$  we simply surround  $P$  by a square  $D$  such that  $D$  fully contains  $P$ . Now, we define a new system, System  $\mathcal{B}$ , which is defined over  $D$ . The arrivals of System  $\mathcal{B}$  within the area  $P$  are identical to those of System  $\mathcal{A}$  while the arrivals in the area  $D - P$  occur with probability 0. Due to the

convexity of this plane, all straight line routes between every two points on  $P$  pass through  $P$ . This implies that the results derived for the gated-greedy and the FCFS policy hold for this system as well. The results derived for the gated-scan policy can be applied to this system by simple modification of the server fixed path.

6.2. *Serving customers on a line.* The problem of serving customers on a *straight line* has several applications. One such application is in the deployment of disk arm movement strategies.

The necessary condition for stability, and the stability of the FCFS and gated-greedy policies, are directly implied by the fact that the straight line can be considered as a special case of the convex plane.

The results for the gated-scan policy change somewhat. The fixed path is now a simple 'elevator' scan of the line, and the local moves of the server are all of duration zero. For this reason one can show that the policy is stable for every  $\rho < 1$ , and in this sense is optimal.

An interesting question is whether the results apply to *non-straight lines*. Note that these are not convex spaces, and thus the above arguments do not hold. Nonetheless, one can map the coordinates of such a line to those of a straight line and thus derive the same results.

6.3. *Serving customers on  $N$ -dimensional convex spaces.* The model discussed in this paper can be generalized to spaces of higher dimensions (three and above). While it may be difficult to describe applications for four-dimensional (or higher-dimensional) spaces, three-dimensional space applications do exist (although, at present, in a science fiction context). Such an application would be a spacecraft that moves between satellites to carry out repairs, and some underwater applications such as the survey of wrecks or repair of oil rigs.

The extension of the square plane results to a three-dimensional cube are quite simple.

1. The necessary condition for the stability of an arbitrary policy and the result for the stability of the FCFS policy can be repeated in a similar manner.

2. The results for the gated-greedy policy can be reproduced with the following modifications:

- the upper bound for  $N_x$  is some constant times  $(d/x)^3$ ;
- $R(N)$  is bounded from above by some constant times  $N^{\frac{2}{3}}$ .

3. The gated-scan policy requires defining the server fixed path over the cube. This can be done by slicing the cube into layers and applying in every layer the fixed path designed in Figure 1 for the square plane. The results then follow in a similar manner.

The generalization to an arbitrary convex three-dimensional space is similar to that of an arbitrary convex plane. The generalization of all these results to  $N$ -dimensional ( $N > 3$  but finite) are done in a similar manner.

6.4. *Non-convex spaces.* Under certain conditions the results of this paper apply to non-convex-space figures as well. Specifically, if a 1–1 continuous mapping  $\phi$  exists from a closed square in  $\mathbb{R}^N$  onto the non-convex figure, also in  $\mathbb{R}^N$ , such that the straight-line segments in the square are mapped to trajectories with bounded length in the non-convex figure, then one could use the image of the gated-greedy discipline on the square to obtain a stable policy (for  $\rho < 1$ ) on the non-convex figure. More precisely, let  $X$  be the set of locations of customers in the non-convex figure at some gating instant. We now place corresponding customers in the square at points  $\{y: \phi(y) \in X\}$ . We then apply the gated-greedy algorithm to the square, and use  $\phi$  to calculate the path of the server in the non-convex figure. However, not every figure in space has such a continuous mapping from the square; for example, Rodin's statue 'The Thinker' does not; in fact, since his elbow touches his knee, he is homomorphic to a torus.

The results concerning the gated-scan policy can however be applied to non-convex figures, provided that for any  $\epsilon$  there exists a finite fixed path such that the distance between any point  $x$  and its fixed path point  $\sigma(x)$  is smaller than  $\epsilon$ .

## References

- [1] ALTMAN E. AND FOSS, S. (1993) Polling on a graph with general arrival and service time. INRIA report No. 1992.
- [2] ALTMAN, E. AND SPIEKSMAN, F. (1994) Geometric ergodicity and moment stability of station times in polling systems. *Stoch. Models*.
- [3] ALTMAN, E., KONSTANTOPOULOS, P. AND LIU, Z. (1992) Stability, monotonicity and invariant quantities in general polling systems. *Queueing Systems, Special Issue on Polling Models*, ed. H. Takagi and O. Boxma, **11**, 35–57.
- [4] ALTMAN, E. FOSS, S. RIEHL, E. AND STIDHAM, S. (1994) Sample path analysis of token rings. In *The Fundamental Role of Teletraffic in the Evolution of Telecommunication Networks, Proc. 14th Internat. Teletraffic Congr. Antibes Juan-les-Pins*, pp. 811–820.
- [5] BACCELLI, F. AND BRÉMAUD, P. (1980). *Palm Probabilities and Stationary Queues*. Lecture Notes in Statistics, Springer-Verlag, Berlin.
- [6] BARTHOLDI, J. J. AND PLATZMAN, L. K. (1988) Heuristic based on spacefilling curves for combinatorial problems in euclidean space. *Management Sci.* **34**, 291–305.
- [7] BEARWOOD, J., HALTON, J. AND HAMMERSLEY, J. (1959) The shortest path through many points. *Proc. Camb. Phil. Soc.* **55**, 299–327.
- [8] BERTSIMAS, D. J. AND VAN RYZIN, G. (1991) A stochastic and dynamic vehicle routing problem in the euclidean plane. *Operat. Res.* **39**, 601–615.
- [9] BOXMA, O. J., LEVY, H. AND YECHIALI, U. (1992) Cyclic reservation schemes for efficient operation of multiple-queue single-server systems. *Ann. Operat. Res.* **35**, 187–208.
- [10] COHEN, J. W. (1982) *The Single Server Queue*. North-Holland, Amsterdam.
- [11] COFFMAN, E. G. JR AND GILBERT, E. N. (1987) Polling and greedy servers on the line. *Queueing Systems* **2**, 115–145.
- [12] COFFMAN, E. G. JR AND STOLYAR, A. (1993) Continuous polling on graphs. *Prob. Eng. Inf. Sci.* **7**, 209–226.
- [13] HAREL, A. AND STULMAN, A. (1994) Polling, greedy, and horizon servers on a circle. *Operat. Res.*
- [14] FRICKER, C. AND JAIBI, M. R. (1992) Monotonicity and stability of periodic polling models. Report FEW 559, Department of Economics, Tilburg University.
- [15] GEORGIADIS, L. AND SZPANKOWSKI, W. (1992) Stability of token passing rings. *Queueing Systems, Special Issue on Polling Models* ed. H. Takagi and O. Boxma, **11**, 7–33.

- [16] KHAMISY, A., ALTMAN, E. AND SIDI, M. (1992) Polling systems with synchronization constraints. *Ann. Operat. Res. Special Issue on Stochastic Modeling of Telecommunication Systems*. ed. P. Nain and K. W. Ross, pp. 231–267.
- [17] KROESE, D. P. AND SCHMIDT, V. (1994) Light-traffic analysis for queues with spatially distributed arrivals.
- [18] LEVY, H., SIDI, M. AND BOXMA, O. J. (1990) Dominance relations in polling systems. *Queueing Systems* **6**, 155–172.
- [19] MEYN, S. P. AND TWEEDIE, R. L. (1992) Stability of Markovian processes I: criteria for discrete time chains. *Adv. Appl. Prob.* **24**, 542–574.
- [20] RESING, J. A. C. (1991) Polling systems and multi-type branching processes. *Report BS-R9128*, C.W.I., Amsterdam.
- [21] TWEEDIE, R. L. (1983) Criteria for rates of convergence of Markov chains, with application to queueing and storage theory. In *Probability, Statistics and Analysis*, ed. J. F. C. Kingman and G. E. H. Reuter, London Math. Society Lecture Notes Series 79, pp. 260–276. Cambridge University Press.
- [22] TWEEDIE, R. L. (1983) The existence of moments for stationary Markov chains. *J. Appl. Prob.* **20**, 191–196.
- [23] ZHDANOV, V. S. AND SAKSONOV, E. A. (1979) Conditions of existence of steady-state modes in cyclic queueing systems. *Autom. Remote Control* **40**, 176–184.