HEWLETT PACKARD

The Cesaro Limit of Departures from Certain ·/GI/1 Queueing Tandems

Tom Mountford[1], Balaji Prabhakar
Basic Research Institute in the
Mathematical Sciences
HP Laboratories Bristol
HPL-BRIMS-96-17
May, 1996

Cesaro limits;
increasing hazard rate
services; couplings

We consider an infinite tandem of independent, identical ./GI/1 queues with mean service rate equal to 1 subjected to stationary and ergodic inputs of rate r < 1. Of some interest in the study of such queueing tandems are the following three inter-related questions: (1) For each r < 1, does there exist a rate r stationary and ergodic process which is an invariant distribution for the queue? (2) For a fixed r, is this invariant distribution unique? (3) When a stationary and ergodic arrival process of rate r <1 is input to the first queue, do the successive departure processes converge in distribution to the invariant distribution (assuming it exists)? For general non-exponential server queues, it is not yet known if invariant distributions exist. However for each r < 1, should one exist, it is known to be unique. This note contributes to the third question when the service time distribution of each queue in the tandem has an {\em increasing hazard rate}. It is shown that when a stationary and ergodic arrival process of rate r <1 is passed through a tandem of such queues, the Cesaro averages of the successive departure processes converge weakly to a limit which is an invariant distribution for the queue.

[1] Department of Mathematics, University of California

THE CESARO LIMIT OF DEPARTURES FROM CERTAIN ·/GI/1 QUEUEING TANDEMS *

TOM MOUNTFORD
Departments of Mathematics
University of California, Los Angeles.
Email: malloy@math.ucla.edu

BALAJI PRABHAKAR
BRIMS
Hewlett-Packard Labs, Bristol.
Email: balaji@hplb.hpl.hp.com

Abstract

We consider an infinite tandem of independent, identical ·/GI/1 queues with mean service rate equal to 1 subjected to stationary and ergodic inputs of rate $\alpha < 1$. Of some interest in the study of such queueing tandems are the following three inter-related questions: (1) For each $\alpha < 1$, does there exist a rate α stationary and ergodic process, $\mathbf{I}_\alpha$, which is an invariant distribution for the queue in the sense that $\mathbf{I}_\alpha \stackrel{d}{=} T(\mathbf{I}_\alpha)$? (Here $T(\mathbf{I}_\alpha)$ is the equilibrium departure process corresponding to an input of $\mathbf{I}_\alpha$). (2) For a fixed α, is this invariant distribution unique? (3) When a stationary and ergodic arrival process of rate $\alpha < 1$ is input to the first queue, do the successive departure processes converge in distribution to the invariant distribution $\mathbf{I}_\alpha$ (assuming it exists)?

For general non-exponential server queues, it is not yet known if invariant distributions exist. However for each $\alpha < 1$, should one exist, it is known to be unique [4, 10]. This note contributes to the third question when the service time distribution of each queue in the tandem has an *increasing hazard rate*. It is shown that when a stationary and ergodic arrival process of rate $\alpha < 1$ is passed through a tandem of such queues, the Cesaro averages of the successive departure processes converge weakly to a limit which is an invariant distribution for the queue.

1 Introduction

In this note we study the effect of passing a subcritical stationary point process through a series of ·/GI/1 queues. We are interested in establishing the distributional convergence of successive departure processes to a limit, assuming that such a limit

*Tom Mountford's research was supported in part by NSF grant DMS9157461, a grant from the Sloan Foundation, and the NSERC.

exists. In [9] coupling arguments were used to establish the Poisson convergence of successive departures from a series of $\cdot/M/1$ queues. Our goal here is to adapt these coupling arguments for non-memoryless services. As opposed to the case of $\cdot/M/1$ queues which are known to possess the Poisson process as the (only) invariant distribution [3, 1], a major problem in the case of non-exponential server queues is that it is not yet known if they possess an invariant distribution. We do not address this question, but rather establish some convergence results which take place when requisite invariant distributions exist. In this sense, this work is somewhat limited. In the remainder of this section we introduce some relevant terminology and describe our result.

Let A^0 be an ergodic stationary point process of rate α strictly less than one. Suppose that it is passed through a $\cdot/GI/1$ queue whose service times have mean one to obtain another stationary ergodic output process A^1 of rate α. That A^1 is stationary and ergodic and of rate α is guaranteed by Loynes' construction (for details of Loynes' construction see, for example, [2]). The output process A^1 can in turn be regarded as an input process for another $\cdot/GI/1$ queue, independent of and identical to the first. As before, we may then obtain a stationary and ergodic process A^2 as output from the second queue. Proceeding thus, we may obtain a series of output processes $A^n, n \in \mathcal{Z}^+$, each stationary and ergodic of rate α. The following is the question we wish to address: what can be said about the point process A^n as $n \longrightarrow \infty$? In the case where the services are i.i.d. exponentials of rate one it was recently shown (see [9]) that as n tends to infinity, A^n converges in distribution to a Poisson process of rate α. We would like to show an analogous result for general service distributions. Basically the method of [9] consisted of inputting a rate α Poisson process P^0 into the "same" queue as A^0 and obtaining a sequence of output processes P^i. By the well known Burke's Theorem, the processes P^i will all be Poisson of rate α. The proof then consisted of showing that as n becomes large, the processes A^n and P^n come closer and closer, and that A^n converged in distribution to a Poisson process.

Say that an input point process A, with law μ, is *invariant* for a $\cdot/GI/1$ queue if the corresponding departure process D is also distributed according to μ. As mentioned above, it is not known whether a general $\cdot/GI/1$ queue with mean service rate equal to one admits an invariant distribution of rate α ($\alpha < 1$). However, it is known that should such an invariant distribution exist, then it must necessarily be unique [4, 10]. This uniqueness of invariant distributions of a given rate is obtained by showing that, as a mapping on the space of stationary point processes, the queueing operator is contractive under a metric $d(.,.)$. That is, for distinct stationary arrival processes, A and B, of rate α, there is a metric $d(.,.)$ such that $d(A,B) > d(T(A),T(B))$, where $T(A)$ (respectively, $T(B)$) is the departure process resulting from an input of A (respectively, B) [4, 10]. The fact that the queueing operator is contractive under this metric also implies that any stationary invariant arrival process of *pathwise rate* α (i.e. a stationary invariant process whose ergodic components are all of rate α) must necessarily be ergodic (see [10]). We are careful to mention this here because later on we establish the convergence of Cesaro means of successive departure processes to a limit which is an invariant distribution for the

queue. By the above discussion, this invariant distribution (assuming it exists) is unique and ergodic.

Attention, in this note, is restricted to the family of service distributions of *increasing hazard rate.* For a random variable X, the hazard rate f (if it exists) is defined by

$$P[X \in (x, x+dx) | X > x] = f(x)dx + o(dx).$$

A random variable (or distribution) has an increasing hazard rate if f is increasing. If a random variable has increasing hazard rate then some exponential moments exist. The family of distributions with increasing hazard rate includes the exponential distribution whose hazard rate is constant on the positive half line. The following is the main result of this paper.

Theorem 1 *Let A^0 be a stationary and ergodic point process of rate α which is passed through a sequence of i.i.d. $\cdot$/GI/1 queues whose service times have increasing hazard rate and have unbounded distributional support. Let the output from the n^{th} queue be A^n with corresponding law μ_n. If the $\cdot$/GI/1 queue admits a rate α invariant distribution, μ, then*

$$\bar{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \mu_i$$

converges in distribution to μ.

Two assumptions are made for the service distribution. The first one, increasing hazard rate, is demanded by our whole approach; the second, unbounded support, is merely a simplification. At the end of the paper we sketch how this extra assumption may be removed.

We can only prove convergence of the Cesaro means because in this case limit points can be shown to be invariant. However, we believe that the original sequence of processes, $\{A^n, n \in \mathcal{Z}^+\}$, converges in distribution to an invariant limit when n tends to infinity.

The proof of the theorem requires a modification of Loynes' construction. We proceed to develop this in the next section.

2 A Modification of Loynes' Construction

In the following, for notational convenience only, we consider simple arrival processes. The standard development of Loynes' construction for a $\cdot$/GI/1 queue can be briefly summarized as follows. Arriving customers are associated with service times chosen in an i.i.d. fashion from some general service time distribution. Without loss of generality, these customers can be assumed to belong to an arrival process A^0 of rate

$\alpha < 1$ and the service times can be assumed to have a mean value equal to one. Letting $A^{0,n}$ denote the (non-stationary) process obtained by suppressing all arrivals before time $-n$, it is then shown that the queue size at time t, X_t^n, resulting from an input of $A^{0,n}$ grows with n along each realization ω of the arrival and service processes. Therefore, $\lim_{n\to\infty} X_t^n = X_t^\infty$ exists. The condition $\alpha < 1$ (arrival rate < service rate) is then used to show that X_t^∞ is an a.s. finite stationary and ergodic process. If $A^{1,n}$ is the output corresponding to an input of $A^{0,n}$, it then easily follows that a stationary limit, A^1, of the non-stationary outputs $A^{1,n}$ exists. See [2] for details.

For the special case of exponential servers, a different method of construction of A^1 from A^0 is used in [7], [1] and [9]. Here, the queue is associated with a rate one Poisson process N^0 representing virtual service times. For $t \in N^0$, a customer is served at time t if and only if the queue is nonempty at time t^-. Defining the service in this way, for an arrival process A^0 of rate $\alpha < 1$ one can construct the output process A^1 and queue process X^∞ as before by considering $A^{0,n}$ and taking the limit as $n \to \infty$. Further details of this procedure may be found in [9].

To study queueing systems with increasing hazard rate, we introduce a further modification to Loynes' construction for memoryless servers. Instead of a simple Poisson process of virtual service times we will derive our queue from a Poisson process N on $R^1 \times R^+$ of Lebesgue intensity. Given a queueing process $X(s)$, $-\infty < s < \infty$, a service will occur at time t if $(t,r) \in N$ for some r with $r \leq f(t)$, where $f(t)$ is the hazard rate of the customer in the queue currently being served. It follows from Markov properties of the Poisson process that for a queueing process $X(t)$, the point process N defines a server with appropriate i.i.d. service times.

The following is a simple large deviations lemma stated without proof.

Lemma 1 *Fix $\beta < 1$. Given an arrival process $\cdots t_{-1} < 0 \leq t_0 < t_1 \cdots$. consider the queueing process X_t^n starting at t_{-n} with a single customer in the queue (the $-n^{th}$ customer), and with service derived from the Poisson process N defined above. If $t_{-n} < -n/\beta$, then the chance that $X_t^n > 0$ for each $t \in [t_{-n}, 0]$ is $\leq Ce^{-nc}$ for some finite, positive C, c.*

It is easy to see that for each t the size of the queue, X_t^n, increases with n. We show that this limit must a.s. be finite.

Lemma 2 *The increasing limit of the processes $X^n, n \in \mathcal{Z}^+$, is a.s. finite as $n \to \infty$.*

Proof Fix $\alpha < \beta < 1$. By the fact that the rate of A is $\alpha < \beta$, there exists M so that for all $n \geq M$, $t_{-n} < -n/\beta$. By Lemma 2.1, for $n \geq M$, the chance that the queue starting with a single customer at time t_{-n} is never empty in $(t_{-n}, 0]$ is $\leq Ce^{-nc}$.

Thus, by the Borel-Cantelli Lemma, there exists M' so that for all $n \geq M'$ the queue started with one customer at time t_{-n} is empty at some time in $(t_{-n}, 0]$. Now, if $X_s^{n+1} = 0$ for $s \in (t_{-n-1}, 0]$, then $X_r^n \geq X_r^{n+1}$ for all $r \in [s, \infty)$. However, by monotonicity in n, $X_r^n \leq X_r^{n+1}$ for $r \in [s, \infty)$. In particular, $X_r^n = X_r^{n+1}$ for all $r \in [0, \infty)$ and, therefore, for all $n \geq M'$, $X_s^n = X_s^{M'}$ for $s \in [0, \infty)$ implying the finiteness of $\lim_{n\to\infty} X_t^n$. ■

Lemma 3 *The limit of the output processes corresponding to the increasing limit of the X^n's exists a.s.*

Proof First suppose that A is ergodic. From the usual Loynes construction it is clear that all we have to do is show that with probability one there exist times tending to $-\infty$ at which $X_t^\infty = \lim_{n\to\infty} X_t^n$ is equal to zero. By ergodicity of (A, N), this will be the case if we can show that the chance that X_t^∞ is zero at a fixed point is strictly positive. By stationarity, this in turn will follow if we can show that with positive probability, on some interval of positive length the limiting queue size is zero. By Lemma 2.2, we can choose L so large that $P[X_0^n < L] > 1/2$ for each n. However in the notation of Lemma 2.1, for m large enough $t_m > m/\beta$. Thus we can apply this lemma to the time interval $[0, t_m]$ (for m large enough), to conclude that the limit of X^n $(= X^\infty)$ is zero on an interval in $(0, m)$ with positive probability. This completes the proof for ergodic A.

If A is of rate α but not ergodic then it can be expressed as a mixture of rate α ergodic point processes. The above argument applied to each of these ergodic processes gives the result for this A. ■

3 A Coupling of Point Processes

In this section we consider two arrival processes A^0 and B^0. A^0 is stationary and ergodic of rate $\alpha < 1$. B^0, also stationary and ergodic of rate α, is the (unique) invariant distribution for our queue (assumed to exist). We suppose that they are served by the same Poisson process N^0 on $R^1 \times R_+$, producing, respectively, A^1 and B^1 as output. These are then fed into a second queue served by an independent Poisson process N^1 (identically distributed to N^0). Proceeding thus, a series of departure processes $\{A^n\}$ and $\{B^n\}$, $n \in \mathcal{Z}^+$, are obtained. We wish to examine how close A^n and B^n become as n tends to infinity. Initially all the customers in the A and B processes are labelled individually. If x is a customer in the A process, then $s_x(n)$ will denote the arrival time of x in the process A^n. Similarly if y is a customer in the B process then $t_y(n)$ will denote the arrival time of y in B^n. Note that in the following, we may switch the ordering of customers in A^i (or B^i) so that if x_1 and x_2 are customers in the A queues, it may be that x_1 arrives at queue i before x_2, but leaves the queue after x_2.

We wish to match up the A and B customers in such a way that for all large n the processes A^n and B^n are "close" to each other in distribution. The aim is that once customers are matched they remain so forever; equivalently, if $x \in A^n$ is matched with $y \in B^n$, then x and y will be matched in (A^m, B^m) for all m greater than n. We now describe how customers are matched up at a queue.

For ease of exposition we introduce a colouring scheme similar to that employed in [9].

If customer $x \in A^n$ is matched with customer $y \in B^n$, then they are both coloured yellow. If customer $x \in A^n$ is unmatched, it is coloured blue. If customer $y \in B^n$ is unmatched, it is coloured red. So red or blue customers may become yellow subsequently, but, we will prove in Proposition 3.1 that once a customer becomes yellow, it will remain so forever. Say that x and y are matched in the pair (A^n, B^n) (or, equivalently, after queue n) if $s_x(n), t_y(n)$ satisfy

$$(*) \qquad s_x(n) = \inf\{s_z(n) \in [t_y(n), t_y^+(n)] : z \text{ is not matched with } y' \neq y\}$$

where $t_y^+(n) = \inf\{t > t_y(n) : t \in B^n\}$. In other words, x and y are matched after queue n if x is the first unmatched A customer to depart from stage n after y, but before the next B customer who departs at time $t_y^+(n)$. This definition is ambiguous only in the trivial case where A^n is identical to B^n.

Observe that x and y may be matched after queue M but that $s_x(M) = t_j(\in B^M) \neq t_y(M)$ or $t_y(M) = s_i(\in A^M) \neq s_x(M)$. That is, matching doesn't imply coincidence of points and vice-versa. This is because in our definition of matching we consider the interval $[t_y(n), t_y^+(n)]$ and not the interval $[t_y(n), t_y^+(n))$.

In order that the coupling argument be successful, it is crucial that once x and y are matched they remain so ever after. It is easy to see that a simple service policy like FIFO does not preserve matchings. For, consider two matched customers x and y belonging respectively to processes A and B at queue n. In particular, this implies that y departs from queue n before x. However, at queue $n+1$, it is quite possible that y is delayed long enough by the presence of other B customers (who will not affect x) so that x departs from this queue before y, thus annulling the matching between x and y. In order to preserve matchings, we will therefore be required to modify the service policy for the A and B customers, sometimes giving priority to yellow customers and sometimes to non-yellow customers. This in itself will not be sufficient to preserve matchings and we will be forced to relabel customers from time to time.

Before describing the priority scheme and the relabelling procedure, we would like to motivate them by giving a small list of things we would want them to achieve.

(a) If $x \in A$ and $y \in B$ are matched after stage n then to preserve $(*)$ we would like to serve y before x at all subsequent stages and once y has been served, we would like to serve x as soon as possible. In this case, we might have to bump

the priority of x (a yellow customer) over that of some other unmatched (blue) A customer.

(b) If x and y are matched after stage n and if y is at the $(n+1)^{th}$ stage while x has yet to arrive, we would like to avoid serving y, if possible.

We now proceed to introduce the priority scheme. Consider customers arriving according to (A^n, B^n) at the $(n+1)^{th}$ queue. These customers are coloured yellow, red or blue depending on whether they are matched when departing from the n^{th} stage or not. We begin by maintaining two imaginary internal queues for each of the A and B customer types. That is, at every stage n, each of the A and B queues will have separate FIFO buffers for yellow and non-yellow customers. Associated with queue A and present time t will be a random variable $\tau_A(t)$ which represents the amount of time that queue A has been non-empty since the last service was rendered. Similarly $\tau_B(t)$ is associated with queue B and time t. If a point (t, r) occurs in the service generating Poisson process N, then a customer will be served at time r at queue A (respectively, B) if and only if $r \leq f(\tau_A(t))$ (respectively, $f(\tau_B(t))$), where f is the hazard rate of the service distribution. The priority scheme and relabelling procedure amount to a rule for deciding when to release a yellow customer over a non-yellow one.

There are essentially three distinct possibilities to address for $(t, r) \in N$:

(1) $f(\tau_B(t)) < r \leq f(\tau_A(t))$: *i.e., there is a service for an A type customer but not for a B type customer.*

A decision as to which A type customer is served needs to be made only if both the internal queues are non-empty. (Equivalently, if there are yellow and non-yellow customers in the A queue.) If both queues are non-empty, then the first yellow customer is released if and only if its matched B partner has already been served. If not, the first blue customer is released. $\diamond$

(2) $f(\tau_A(t)) < r \leq f(\tau_B(t))$: *i.e., there is a service for a B customer but not for an A customer.*

Again a decision about which type of B customer is to be released needs to be made only if both internal queues are non-empty. If we have a choice between a red and a yellow customer, then the first yellow customer is released if and only if its matched A partner is presently in the A queue. Otherwise, the first red customer is released. $\diamond$

(3) $r \leq f(\tau_A(t)), f(\tau_B(t))$: *i.e., there is a service for both an A customer and a B customer.*

Consider the B queue first. If there are both yellow and red customers, then release the first yellow customer only if its partner is also in the A queue. If the

partner has not yet arrived to the A queue, then release the first red customer. If there are only yellow (red) customers present, then the first yellow (respectively, red) customer has to be released, since service is non-idling.

Consider the A queue next. If there are any yellow customers present, then release the first of these. If there are no yellow customers, then release the first blue customer. $\diamond$

These are the rules for re-prioritizing customers and we will now describe the relabelling procedure. But first, we briefly discuss the need to do this.

Consider Rule (2) above. It is quite possible that the only B customer present is a yellow customer, y, while queue A has only blue customers. This means that the partner of y, x, is yet to arrive at queue A. Now, it is possible that between the departure of y and the arrival of x there is service at queue A. Should this happen one of the blue customers, z say, is let go and this leads to an annulment of the matching between x and y, since x will no longer be the first A customer departing from stage $n+1$ after y.

A similar situation is encountered under Rule (3) when we are forced to release a yellow B customer, say y, and a blue A customer, say z, simultaneously while x, the partner of y, has yet to arrive.

In either of the above cases, we simply swap the labels of customers x and z to ensure that the first A customer to leave queue $n+1$ in time $[t_y(n+1), \infty)$ is the partner of y, i.e. x.

We are now ready to prove the following proposition which establishes that under the above-stated rules for customer re-prioritizing and relabelling, the matchings of customers are preserved forever.

Proposition 1 For each n, if customers x and y belonging, respectively, to processes A and B are matched after queue n, then they will continue to be matched after queue $n+1$.

Proof We wish to show that all matched pairs of customers in (A^n, B^n) are also matched in (A^{n+1}, B^{n+1}). By the construction in Section 2, it is sufficient to show this for processes begun at time $-N$, with empty queues. We may disregard the arrival of the first yellow A customer if its partner arrived before time $-N$. Look at the first time that there is a departure so that the relation $(*)$ fails for some previously matched customers x and y. This could occur in three ways. Either $s_x(n) < t_y(n)$, $s_x(n) > t_y^+(n)$ or there are intervening blue customers between y and x. The last cannot occur because of relabelling so we concentrate on the first two.

Suppose that it is the first and, hence, that x leaves queue n before y. Now x arrived at the A queue after y arrived at the B queue. So y must be queueing while

x departs. By Rules (1) and (3), when customer x is served there must be no blue customers present and x must be the longest waiting yellow customer in the A queue. By Rule (3), there cannot be a B customer served at the same time as x is served since this would require that either y be this B customer, or that (x, y) not be the first matched pair who fail to satisfy $(*)$. Therefore, the three facts (a) x is served alone, (b) when x is served there are no blues, and (c) y arrived before x, together imply that the last service of any kind before x is served must be that of a B customer but not of an A customer (recall that service has increasing hazard rates).

Now, if this last service occurred before the arrival of x, then at this point the A queue must have been empty and, since y arrives before x, increasing hazard rate properties of the server imply that a service must take place at queue B before or with the next service at the queue A. On the other hand if it occurred after the arrival of x, then by Rule (2) y must be served since it is the first yellow in the B queue (using the fact that (x, y) is the first pair who become unmatched). Either way there is a contradiction.

Next suppose that $s_x(n) > t_y^+(n)$. Let w be the B customer who departs after y at time $t_y^+(n)$. Now, customer y is either served alone or simultaneously with an A customer. Since the argument is similar in both cases, we only treat the case where y is served alone. Suppose first that x is present in the A queue when y is served. Because of increasing hazard rate properties of the server, the next customer served will be of type A (possibly simultaneously with a B customer). Since we assume that (x, y) is the first pair for whom things go wrong, by Rule (1), this next A customer must be x. This contradicts $s_x(n) > t_y^+(n)$. On the other hand, if x is not present when y is served then, by Rule (2), at time $t_y(n)$ there must be no red customers in the B queue. By property $(*)$ there are also no other yellow customers in the B queue. Thus when x arrives (possibly with a B customer) the B queue is empty. Thus the next service after $t_y(n)$ must be of an A customer (possibly simultaneously with a B customer). By Rules (1) and (3) and the assumption that (x, y) is the first pair to go wrong, this A customer must be x. Again, this contradicts $s_x(n) > t_y^+(n)$. ■

Similar arguments establish

Proposition 2 *Suppose that after queue n customer x is coloured blue, customer y is coloured red and $t_y(n) < s_x(n)$. If $s_x(n+1) < t_y(n+1)$ (that is, x overtakes y at queue n), then x must be coloured yellow at all queues after $n+1$.*

Sketch of Proof: Suppose that blue customer x arrives at queue $n+1$ after red customer y, but leaves strictly before y. Let $s_x^-(n+1)$ be the time of the A-departure from queue $n+1$ immediately prior to $s_x(n+1)$. Then either queue B is non-empty throughout the interval $(s_x^-(n+1), s_x(n+1)]$ or $s_x^-(n+1) < t_y(n) < s_x(n)$ and queue A is empty in the interval $(s_x^-(n+1), s_x(n))$. Either way a (necessarily red) customer must be released from queue B in the interval $(s_x^-(n+1), s_x(n+1)]$.

4 Proof of eventual matching of customers

For ease of exposition, we will assume throughout this section that the support of the service times is unbounded. With appropriate non-trivial modifications, the methods can be extended to the case of bounded service times using techniques similar to those in [10]. The remark at the end of the paper provides an idea of some of the steps involved in implementing the changes. We are now ready to prove the following proposition.

Proposition 3 *If B^0 is an invariant distribution of rate α and A^0 is stationary and ergodic of rate α, then the densities of matched customers in A^n and B^n must increase to α as n increases to ∞.*

Proof The argument is essentially an adaptation of that given in [6]. We argue by contradiction and suppose that the density of matched customers increases to a value strictly less than α. Then we must have customers who remain red or blue forever. Let us call such customers ever-reds and ever-blues.

By the preservation of order among blues and reds established in Proposition 3.2, if customer x is an ever-blue, customer y is an ever-red, and $s_x(0) < t_y(0)$, then, for each n, it must be the case that $s_x(n) < t_y(n)$. This enables us to talk of intervals of ever-reds and ever-blues. We say customers $y_1, y_2, ..., y_r$ with $t_{y_1}(0) < t_{y_2}(0) < ...t_{y_r}(0)$ form an interval of ever-reds if

(i) For all i, y_i is an ever-red.

(ii) There is no ever-blue x with $t_{y_i}(0) \leq s_x(0) \leq t_{y_j}(0)$.

(iii) The interval (y_1, y_r) is not contained in a strictly larger subset with properties (i) and (ii).

We similarly define intervals of ever-blues.

Given such an interval (y_1, y_r) of ever-reds with $t_{y_1}(0) < t_{y_r}(0)$, call y_1 a left ever-red and y_r a right ever-red. A similar labelling procedure applies for ever-blues.

Now, consider the process of ever-reds and ever-blues. By the stationarity of (A^0, B^0), and the translation invariant nature of the queueing operation, the process of ever-reds and ever-blues must be stationary. However, a priori these need not be ergodic and hence the density of the ever-reds (and similarly of the ever-blues) must be considered a random quantity R where $E(R) < \alpha$. However the following fact follows from ergodicity of B^n. The (non-random) density of red customers in B^n must decrease to $E(R)$ as n goes to infinity. Similarly for blue customers in A^n. Thus R is non-random.

Therefore initially there coexist, with probability one, ever-blue and ever-red customers, both of density $R > 0$. Consider the points of $A^0 \cup B^0$ that represent customers that never become yellow as an alternation of intervals of ever-blue customers with intervals of ever-red customers. By the order preservation property shown in Proposition 3.2, these intervals will have their "orderings" preserved. This and the fact that ever-reds and ever-blues coexist a.s. imply that the processes of left ever-red and left ever-blue customers as defined above exist. Now the process of left ever-reds is stationary (for the same reason that the ever-reds are stationary) and thus possesses a (possibly random) density. Denote this density by the random variable L_R and note that it must be true that $E(L_R) > 0$. Therefore for some $\epsilon > 0$, $P(L_R > \epsilon) > \epsilon$. Now, $L_R = L_R^l + L_R^g$, where L_R^l is the density of left ever-reds y_1 such that there is another left ever-red $y_2 \in (y_1, y_1 + 2/\epsilon]$ and L_R^g is the density of the remaining left ever-reds for whom the next left ever-red is at a distance greater than $2/\epsilon$. Since, by definition, $L_R^g \leq \epsilon/2$ P a.s., this means $L_R^l \geq \epsilon/2$ whenever $L_R > \epsilon$.

A vital observation originating in Ekhaus and Gray [6] is that between two left ever-reds there must be a left ever-blue, by definition of intervals of ever-reds/ ever-blues. We deduce from this observation and the previous paragraph that the density of ever-reds y in B^n such that there is an ever-blue x with $s_x(n) \in (t_y(n), t_y(n) + 2/\epsilon]$ must be at least $\epsilon/2$ with probability ϵ. Now ever-blues and ever-reds are distinguished customers, identified by looking into the future, however they must, respectively, be blue and red in A^n and B^n for any n. It now follows from the above that the density of reds y in B^n such that there is a blue x with $s_x(n) \in (t_y(n), t_y(n) + 2/\epsilon]$ must be at least $\epsilon/2$ with probability ϵ. Appealing now to the ergodicity of the pair (A^n, B^n) it is easy to see that this density must, in fact, be $\geq \epsilon/2$ with probability 1.

By the stability of the queues and the S.L.L.N. every customer y in B^n has a strictly positive chance of arriving at an empty queue. Denote this probability by $p_n(y)$. Let us fix $\delta > 0$ so that (for all n by the invariance of B^0), the density of customers y in B^n with $p_n(y) < \delta$ is less than $\epsilon/2$. Therefore it follows (for every n) that with density at least $\epsilon/2 - \epsilon/4 = \epsilon/4$ there are customers $y \in B^n$ so that

(i) there exists a blue customer in $[t_y(n), t_y(n) + 2/\epsilon]$

(ii) $p_y(n) > \delta$.

Suppose now that after such a customer y arrives, it is not served within time $2/\epsilon$ (this event has strictly positive probability). Then it is easy to see that when y leaves, it must be matched to a blue customer.

Thus the density of matched customers in $A^{n+1}-$ the density of matched customers in $A^n \geq (\epsilon/4)\,\delta\,(1 - F(2/\epsilon))$, where F is the c.d.f. of the service distribution. By our unboundedness assumption for the support of service times, this is a positive number. Therefore the density of matched customers in A^n (and B^n) must grow at least linearly with n. This contradicts the fact that the density of A^n, and hence that of the matched customers of A^n, is bounded by α for each n. ■

The proof of Theorem 1.1 is now completed in the next section.

5 Proof of Theorem 1.1

We use a simple uniformity property for Loynes construction.

Lemma 4 *Consider a stationary and ergodic arrival process V^0 of rate $\alpha < 1$ for a $\cdot/GI/1$ queue whose service times have an increasing hazard rate and are of mean 1. Suppose the stationary and ergodic equilibrium output, V^1, has the property that for all $x \geq x_0$,*

$$P\left[\frac{1}{x}V^1(-x,0) > \alpha + \epsilon\right] < \epsilon$$

where $\alpha + 2\epsilon < 1$, then

$$P\left[V^{n,1} = V^1 \text{ on } [0,\infty)\right] \geq 1 - \epsilon - e^{-nc(\epsilon)}$$

for some $c(\epsilon)$ strictly positive. Here $V^{n,1}$ is the output obtained from $V^{n,0}$ and $V^{n,0}$ is as in Section 2.

Proof As has already been noted, if X^∞ is the limiting queueing process for V^0 and X^n is the queueing process for $V^{n,0}$, then $X^\infty \geq X^n$. From this and the Markov property for our Poisson process, it follows that if $X_t^\infty = 0$ for some $t \in (-n,0)$, then $V^{n,1} = V^1$ on $[0,\infty)$. But because the hazard rate of the service times is increasing, if $X^\infty > 0$ in $(-n,0)$, then it must be the case that the distribution of the number of points of V^1 in $(-n,0)$ is stochastically greater than the distribution of the number of service renewals in $(-n,0)$. Thus by standard large deviations results (recall service times have some exponential moments) $P\left[V^1(-n,0) \leq n(\alpha+\epsilon),\ X^\infty > 0 \text{ on } (-n,0)\right] < e^{-c(\epsilon)n}$. ■

Corollary 1 *Let $A^{r+1,k}$ be the point process obtained by suppressing all points in A^r that occur before time $-k$ and then passing the resulting point process through the $\cdot/GI/1$ queue generated by Poisson process N^r. For each $\epsilon > 0$ sufficiently small, there exists k sufficiently large so that for all r large enough*

$$P\left[A^{r+1,k} = A^{r+1} \text{ on } (0,\infty)\right] > 1 - \epsilon.$$

Proof Fix $\epsilon > 0$. By ergodicity of the B point processes, there exists x_0 so that for all $x > x_0$ (and for all r, since the B processes are identically distributed for all r),

$P[B^r(-x,0) > x(\alpha+\epsilon/2)] < \epsilon/2$. On any interval, the number of matched points in B^r can differ from the number of matched points in A^r by at most one. Therefore

$$\frac{A^{r+1}(-x,0)}{x} \leq \frac{B^{r+1}(-x,0)}{x} + \frac{C^{r+1}(-x,0)}{x} + \frac{1}{x}$$

where C^{r+1} denotes the unmatched points in A^{r+1}. Now by Proposition 4.1, the density of unmatched points in A^n tends to zero as n tends to infinity. Therefore

$$P\left[\frac{A^{r+1}(-x,0)}{x} > \alpha + 2\epsilon/3\right] < 2\epsilon/3$$

for r sufficiently large if x is greater than x_0. The statement of the Corollary now follows from Lemma 5.1. ■

Proof of Theorem 1.1 Proposition 4.1 implies that the sequence of measures $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\mu_i$ is tight and that any limit point must have density α. It remains to show that these measures converge to μ, the invariant measure of density α. By the uniqueness result of [4], it is sufficient to show that any limit point is an invariant distribution for the $\cdot$/GI/1 queue. In the following T stands for the queueing operator, i.e. $T(\mu_i) = \mu_{i+1}$.

So we must show that if $\bar{\mu}_{n_i} \to \nu$ in distribution, then $< \nu, f > \; = \; < T\nu, f >$ for every bounded continuous function f of compact support. See, for example, [5]. Or equivalently, $< \nu, f > \; = \; < \nu, T^*f >$ for every bounded continuous function f of compact support, where T^* is the adjoint of T.

We suppose without loss of generality that f has support in (0,∞). $\bar{\mu}_{n_i} \to \nu$ weakly, implies that $< \bar{\mu}_{n_i}, f > \; \to \; < \nu, f >$. Also

$$< \bar{\mu}_{n_i}, T^*f > \; = \; \frac{1}{n_i}\sum_{j=1}^{n_i} < \mu_j, T^*f > \; = \; \frac{1}{n_i}\sum_{j=1}^{n_i} < T\mu_j, f > \; = \; \frac{1}{n_i}\sum_{j=1}^{n_i} < \mu_{j+1}, f >$$

and the last equation implies that

$$< \bar{\mu}_{n_i}, T^*f > \; = \; < \bar{\mu}_{n_i}, f > + \frac{1}{n_i}(< \mu_{n_i+1}, f > - < \mu_1, f >) \quad \longrightarrow \quad < \nu, f > .$$

However it is not immediate that as i tends to infinity, $< \bar{\mu}_{n_i}, T^*f > \; \to \; < \nu, T^*f >$ since the function T^*f is not necessarily continuous on the set of discrete measures. We can address this technical point by introducing continuous functions $T^{*,k}f(N) = T^*f(N^{-k})$ where N^{-k} is obtained from N as follows. All arrivals of N before $-(k+1)$ are removed with probability one, customers who arrive at $s \in [-(k+1), -k]$ are removed with probability $k + 1 + s$, and all arrivals after time $-k$ are kept. This

yields a continuous function as the influence of points of N near the cutoff goes to zero. By Corollary 5.2, for every $\epsilon > 0$ we can find a k so that

$$\limsup \left| < \bar{\mu}_{n_i}, T^* f > - < \bar{\mu}_{n_i}, T^{*,k} f > \right| < \epsilon$$

and

$$\left| < \nu, T^* f > - < \nu, T^{*,k} f > \right| < \epsilon.$$

As convergence in distribution of μ_{n_i} to ν implies that for each k

$$< \bar{\mu}_{n_i}, T^{*,k} f > \; \rightarrow \; < \nu, T^{*,k} f >,$$

we now have that $< \bar{\mu}_{n_i}, T^* f > \; \rightarrow \; < \nu, T^* f >$ and, consequently, that $< \nu, f > \; = \; < T\nu, f >$. This proves Theorem 1.1. □

Remark. Extension to the case of bounded service times: The main point at which the unboundedness assumption on the service time distributions is invoked is in the last paragraph of Section 4. Note that by order preservation properties of ever-reds and ever-blues, we are guaranteed a minimum density (at least $\epsilon/4$) of blue customers who are within a distance of $2/\epsilon$ of red customers at every stage n. Now, given that the service times are unbounded, this means that with probability $1 - F(2/\epsilon)$ the reds may be held at a stage for at least $2/\epsilon$ time units, causing a minimum density of customer matchings at each stage and generating the desired contradiction.

Obviously, this argument fails when the service times are bounded and modifications are necessary. The key idea is to realize that although red and blue customers are not guaranteed to match at one stage (because of bounded service times), they can be forced to match over several stages. This is done by giving a red customer higher service times than a blue customer behind it over several stages, thus causing the blue to "catch up" with the red and couple with it. Since the blues in question are within $2/\epsilon$, and the service times are i.i.d., the chance of accomplishing this within a certain fixed number, say S, of stages is strictly positive. This, then, is a description (*sans* details) of how one might extend the result of the paper to the case of $\cdot$/GI/1 queues with bounded service times.

References

[1] V. Anantharam (1993). Uniqueness of stationary ergodic fixed point for a $\cdot$/M/k node. *Annals of Applied Probability*, **3**, 1, 154-173.

[2] F. Baccelli and P. Brémaud (1987). *Palm Probabilities and Stationary Queues.* Lecture Notes in Statistics **41**, Springer-Verlag, New York.

[3] P.J. Burke (1956). The output of a queueing system. *Operations Research*, **4**, 699-704.

[4] C.S. Chang (1994). On the Input-Output Map of a G/G/1 Queue. *Journal of Applied Probability*, **31**, 4, 1128-1133.

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

[6] M. Ekhaus and L. Gray (1993). A Strong Law for the Motion of Interfaces in Particle Systems. *In preparation.*

[7] T.M. Liggett and T. Shiga. A note on the departure process of an infinite series of $\cdot/M/1$ nodes. *Unpublished manuscript.*

[8] R.M. Loynes (1962). The stability of a queue with non-independent inter-arrival and service times. *Proc. Camb. Philos. Soc.*, **58**, 497-520, 1962.

[9] T.S. Mountford and B. Prabhakar (1995). On the weak convergence of departures from an infinite sequence of $\cdot/M/1$ queues. *Annals of Applied Probability*, **5**, 1, 121-127.

[10] B. Prabhakar (1995): Uniqueness of fixed points for the $\cdot/GI/1$ queueing operator. *In preparation.*