ELSEVIER

Systems & Control Letters 28 (1996) 239–245

SYSTEMS & CONTROL LETTERS

On a singular feature of critical G/M/1 queues[1]

Balaji Prabhakar, Nicholas Bambos*

Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA 90024, USA

Received 25 August 1995; revised 6 December 1995

Abstract

In this paper, we consider a critically loaded G/M/1 queue and contrast its transient behaviour with the transient behaviour of stable (or unstable) G/M/1 queues. We show that the departure process from a critical G/M/1 queue *converges weakly* to a Poisson process. However, as opposed to the stable (or unstable) case, we show that the departure process of a critical GI/M/1 queue *does not couple in finite time* with a Poisson process (even though it converges weakly to one). Thus, as the traffic intensity (ratio of arrival to service rates), ρ, ranges over $(0, \infty)$, the point $\rho = 1$ represents a *singularity* with regard to the convergence mode of the departure process.

1. Introduction

The transient and equilibrium behaviour of stable (service rate > arrival rate) queues with Poisson inputs has been studied using Markov chain theory and reversibility [3, 5, 6, 9]. The method of Loynes [7] provides an understanding of their behaviour in the stationary and ergodic context.

A key concept in the approach of Loynes and, indeed, in the general stability theory of queueing systems is the notion of "finite time coupling" between processes (see [1, 7]). Briefly, this has the following meaning. Suppose that $X^0(t)$ is the queue-size of a G/G/1 queue at time $t > 0$ starting with an empty queue at time 0 and that $\mathbf{D}^0$ is the corresponding departure process. Then, provided the queue is stable (i.e. arrival rate < service rate), Loynes [7] has shown that $X^0(\cdot)$ *couples in finite time* with a stationary and ergodic process $X(\cdot)$ and that $\mathbf{D}^0$ *couples in finite time* with a stationary and ergodic process $\mathbf{D}$. That is, there exists a finite random time τ such that a.s. for all $t > \tau$, $X^0(t) = X(t)$, and the paths of $\mathbf{D}^0$ and $\mathbf{D}$ coincide. The processes $X(\cdot)$ and $\mathbf{D}$ are thought of as the equilibrium queue-size and departure processes from the G/G/1 queue. As a consequence of the finite-time coupling property one then deduces that the processes $X(\cdot)$ and $\mathbf{D}$ are unique, i.e. a stable G/G/1 queue admits a unique stationary regime. Finite time coupling is also used to establish the stability of acyclic *networks* of ·/G/1 queues and in demonstrating the existence and uniqueness of stationary operating regimes for such networks [1, 7].

The purpose of this paper is to investigate the finite time coupling property in the context of critically stable G/M/1 queues. It will be shown that in this case the property collapses, leading to a singular behaviour. A variation of Loynes' argument is first used to show that for critical G/M/1 queues: (1) the queue-size process, $X^0(t)$, goes to infinity in distribution as $t \to \infty$ although it visits every state infinitely often (Lemma 2), and (2) the departure process $\mathbf{D}^0$ converges weakly (in distribution) to a Poisson process (Theorem 1). Although these two facts are in keeping with intuition developed from classical queueing analysis, establishing them has proved non-trivial.

[1] Research supported by NSF grants DMI-9216034, NCR-9116268, and a National Young Investigator Award NCR-9258507.

* Corresponding author.

PII S0167-6911(96)00019-9

This is because non-renewal type arrivals make it impossible to use elementary Markov chain theory, while criticality makes it difficult to use rate-based arguments as in the case of stable or unstable queues. This has led to the development of some non-standard techniques (like those in Lemma 1 and Theorem 2) which seem to be of interest in their own right.

The above developments lead to the following interesting result which sets the critical ·/M/1 queue apart from the stable and unstable ones: (3) the departure process, $\mathbf{D}^0$, from a class of critical GI/M/1 queues *does not couple in finite time* with a Poisson process even though it converges weakly to one (Theorem 2). Hence, the convergence mode of the output from a GI/M/1 node changes from strong (finite-time coupling) to weak convergence, depending on whether the value of the traffic intensity (arrival rate/service rate) lies in the set $(0,1)\cup(1,\infty)$ or equals 1.

Our study of this special feature was initiated during the analysis of certain types of parallel processing and manufacturing systems: those involving synchronizations and pipelining (see [8]). We believe that beyond its theoretical interest, a study of this singular behaviour will be useful in other practical situations, where critically (heavily) loaded queues, possibly with finite buffers and blocking, are approximated by their infinite buffer counterparts for the purpose of analysis.

1.1. Notation

Consider a first-come-first-served, ·/M/1 queueing node with mean service time equal to 1. Let the arrival process, $\mathbf{A}$, be given by

$$\mathbf{A} = \sum_{n=-\infty}^{\infty} \delta_{t_n^{\mathrm{a}}}, \tag{1}$$

where $\cdots < t_{-1}^{\mathrm{a}} < t_0^{\mathrm{a}} \leqslant 0 < t_1^{\mathrm{a}} < \cdots < t_n^{\mathrm{a}} < t_{n+1}^{\mathrm{a}} < \cdots$ pathwise, and δ_x is the point mass at x. t_n^{a} is the arrival time of the nth job or customer to the queueing node. We assume that $\mathbf{A}$ is a stationary and ergodic process with respect to the transformation (time shift)

$$\Theta_t\mathbf{A} = \sum_{n=-\infty}^{\infty} \delta_{(t_n^{\mathrm{a}}-t)}.$$

Let $N_{\mathrm{a}}(t_0,t]$ be the number of arrivals to the node in $(t_0,t]$. We abbreviate $N_{\mathrm{a}}(0,t]$ as $N_{\mathrm{a}}(t)$. The mean arrival rate λ_{A} is then equal to $E(N_{\mathrm{a}}(1))$. Recall that a ·/M/1 queueing node with mean service rate equal to 1 and arrival rate equal to λ_{A} is said to be stable, critical or unstable depending on whether $\lambda_{\mathrm{A}} < 1$, $\lambda_{\mathrm{A}} = 1$ or $\lambda_{\mathrm{A}} > 1$. In this paper we will be concerned exclusively with critical ·/M/1 nodes.

Let $\mathbf{S}$ be the 'service process' (or virtual departure process) of the ·/M/1 queueing node. $\mathbf{S}$ is a Poisson process of rate 1 that is independent of $\mathbf{A}$. Let $\mathbf{S}$ be given by

$$\mathbf{S} = \sum_{n=-\infty}^{\infty} \delta_{t_n^{\mathrm{s}}}, \tag{2}$$

where $\cdots < t_{-1}^{\mathrm{s}} < t_0^{\mathrm{s}} \leqslant 0 < t_1^{\mathrm{s}} < \cdots < t_n^{\mathrm{s}} < t_{n+1}^{\mathrm{s}} < \cdots$ pathwise, and t_n^{s} is the arrival time of the nth service token. Let $N_{\mathrm{s}}(t_0,t]$ be the number of service tokens arriving to the node in $(t_0,t]$, and let $N_{\mathrm{s}}(t) = N_{\mathrm{s}}(0,t]$. Note that $E(N_{\mathrm{s}}(1)) = 1$. All random quantities are assumed to be defined on some common probability space $(\Omega,\mathscr{F},P)$.

Given any two realizations of $\mathbf{A}$ and $\mathbf{S}$, using Loynes' construction (see [1] or [7]), we obtain for $s < t$ the queue size process $X^s(t)$ which represents the number of unserved customers in the queue at time t if the queue started processing arrivals at time s, having been empty before that time. Let $\mathbf{D}^s$ be the corresponding departure process. In this letter we are interested in studying the transient behaviour of critical ·/M/1 nodes and contrasting it with the transient behaviour of stable and unstable ·/M/1 nodes. The issues of interest to us are the characteristics of the queue-size process $X^0(\cdot)$ and the associated departure process $\mathbf{D}^0$.

2. The output of critical G/M/1 queues

The main results of this section are establishing the weak convergence of the output of a critical G/M/1 queue to a Poisson process (Theorem 1) and showing that the output of a critical GI/M/1 node does not couple in finite time with a Poisson process (Theorem 2). Lemma 2 establishes the null-recurrence-type behaviour of the queue-size process of a critical G/M/1 node, to be used in the proofs of Theorems 1 and 2.

Given $\mathbf{A}$ and $\mathbf{S}$ we construct for $s \in \mathscr{R}$, the departure process $\mathbf{D}^s$ by ignoring all points of $\mathbf{A}$ which are less than s. Thus, $\mathbf{D}^s$ is the departure process from the queue if it were started empty at time s and is processing arrivals since that time. The corresponding queue-size process $X^s(t)$ is a non-negative integer-valued process with almost surely right continuous paths which are constant outside points in **A** and **S** and varying at points of **A** and **S** as follows:

$$X^s(t) = \begin{cases} 0 & \text{if } t<s \\ X^s(t^-)+1 & \text{for } t\geqslant s,\ t\in \mathbf{A} \\ X^s(t^-)-1 & \text{for } t\geqslant s,\ t\in \mathbf{S} \\ & \quad \text{and } X^s(t^-)>0. \end{cases} \tag{3}$$

Of course by the independence of **A** and **S** we need not consider points in both **A** and **S**. The departure process $\mathbf{D}^s$ can then be defined as the points $\{t \geqslant s : t\in \mathbf{S} \text{ and } X^s(t^-)>0\}$. Specifically,

$$\mathbf{D}^0 = \mathbf{S}1_{\{X^0(t^-)>0\}}. \tag{4}$$

Reasoning in a manner similar to the one in Loynes' construction [1] it is not hard to see that $X^s(t)$ increases as s decreases. Fig. 1 illustrates the situation for some fixed but arbitrary realizations of **A** and **S**. It is seen that for $u<s$, $X^u(t)>X^s(t)$.

A quick explanation for this can be formulated along the following lines. Notice that $X^u(s)\geqslant 0 = X^s(s)$. All upward jumps of $X^u(t)$ and $X^s(t)$ coincide for $t>s$ (same arrivals); and a downward jump of $X^u(\cdot)$ at some time t implies a downward jump of $X^s(\cdot)$ at t, unless $X^s(t^-)=0$. These facts ensure that $X^u(t)$ is bigger than or equal to $X^s(t)$ for all $t>s$. Thus, $\lim_{s\to-\infty} X^s(t) = X(t)$ exists pathwise, *whatever* the value of the arrival rate is. We also recall the following additional details of Loynes' construction, to be used below.

Fact 1. *$X^s(t) \overset{\mathscr{D}}{=} X^{s+p}(t+p)$ for all $p\in\mathscr{R}$; in particular, $X^{-t}(0) \overset{\mathscr{D}}{=} X^0(t)$. This follows from the joint stationarity of the processes* **A** *and* **S**.

Fact 2. *If $\lim_{s\to-\infty} X^s(t) = X(t) < \infty$ a.s., then $X(t)$ is a stationary and ergodic process.*

Before we proceed to our first lemma, we recall the following standard notation from [1] regarding the structure of the canonical space of point processes (the space of realizations). Let **M** be the set of all counting measures, m, on $\mathscr{R}$. Endow **M** with the sigma field $\mathscr{M}$ generated by functions $m \to m(C)$, where C is a Borel subset of $\mathscr{R}$. The pair $(\mathbf{M},\mathscr{M})$ is known as the *canonical space of point processes*; and a point process is thought of as a measurable mapping from $(\Omega,\mathscr{F},P)$ to $(\mathbf{M},\mathscr{M})$. For issues regarding the joint behaviour of a pair of point processes, one works with the product space $(\mathbf{M}\times\mathbf{M},\mathscr{M}\times\mathscr{M})$ and the corresponding time shift $\Theta_s^2 = \Theta_s \times \Theta_s$.

Lemma 1. *Let* **A** *and* **S** *be two jointly stationary and ergodic point processes of rate* 1. *Let the process $\mathbf{B}(\omega)\subset\mathbf{A}(\omega)$ be the set of those points $t_b\in\mathbf{A}(\omega)$ such that $N_a(t_b,t]\geqslant N_s(t_b,t]+1$ $\forall t>t_b$. Then* **B** *is a stationary and ergodic process with mean rate $\lambda_B = E(N_b(1)) = 0$. In particular, the event $\{\omega\colon \mathbf{B}(\omega)(\mathscr{R})>0\}$ has zero probability.*

Proof. Note that **B** is the process that consists of the *bad* points of **A** – starting from any of these points t_b we have at least one more point of **A** than of **S** for all time bigger than t_b. Given the joint stationarity and ergodicity of **A** and **S**, it is easy to see that **B** is stationary and ergodic with respect to Θ_s. Indeed, if $G\in\mathscr{M}$ and $H\in\mathscr{M}\times\mathscr{M}$ are such that

$$\mathbf{B}(\omega)\in G \Leftrightarrow (\mathbf{A}(\omega),\mathbf{S}(\omega))\in H,$$

then

$$P(\mathbf{B}\in G) = P((\mathbf{A},\mathbf{S})\in H) = P(\Theta_s^2(\mathbf{A},\mathbf{S})\in H)$$
$$= P(\Theta_s(\mathbf{B})\in G)$$

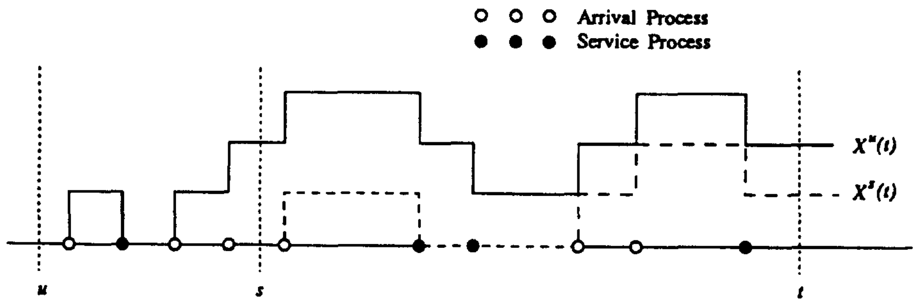

Fig. 1. Loynes' construction.

verifies stationarity of $\mathbf{B}$. Ergodicity follows from the fact that if G is an invariant event for the process $\mathbf{B}$ under Θ_s then H is an invariant event for $(\mathbf{A}, \mathbf{S})$ under Θ_s^2. The Θ_s^2 ergodicity of $(\mathbf{A}, \mathbf{S})$ implies that $P(H) = 0$ or 1; and this implies $P(G) = 0$ or 1.

The process $\mathbf{B}$ thus has a well-defined average rate function which is almost surely equal to the constant $\lambda_B = E(N_b(1))$. Let $\mathbf{B}$ have the representation $\mathbf{B} = \sum_{n=-\infty}^{\infty} \delta_{t_n^b}$, where $\cdots < t_{-1}^b < t_0^b \leqslant 0 < t_1^b < \cdots < t_n^b < t_{n+1}^b < \cdots$.

By definition of the process $\mathbf{B}$, $N_a(t_1^b, t] \geqslant N_s(t_1^b, t] + 1 \ \forall t > t_1^b$. But, more importantly, notice that $N_a(t_1^b, t] \geqslant N_s(t_1^b, t] + n$ if $t_n^b \in (t_1^b, t]$. That is, the difference between $N_a(t_1^b, t]$ and $N_s(t_1^b, t]$ grows by 1 as soon as the time variable t encounters (and exceeds) a point of the process $\mathbf{B}$. Thus, $N_a(t_1^b, t] \geqslant N_s(t_1^b, t] + N_b(t_1^b, t]$ *with probability one*. Dividing by t and taking the limit as $t \to \infty$ and appealing to the ergodic theorem, we get $1 \geqslant 1 + \lambda_B$. This implies that $\lambda_B = 0$. Hence, almost surely the process $\mathbf{B}$ does not exist. This implies that the event $\{\omega\colon \mathbf{B}(\omega)(\mathscr{R}) > 0\}$ has zero probability. □

Our next lemma establishes the null-recurrent type behaviour of the queue-size process of a critical G/M/1 queue. Indeed, this fact would be immediate if we were looking at critical GI/M/1 queues (and not the bigger class of G/M/1 queues); for, then the embedded Markov chain obtained by looking at the queue-size at the instant of an arrival is a null-recurrent one. In the following lemma and throughout the rest of the paper we set $Q(t) = X^0(t)$.

Lemma 2. *For a critical G/M/1 node with stationary and ergodic arrival process* $\mathbf{A}$, *the following statements are true*:

(1) *The queue-size process* $X(t)$ *given by* $X(t) = \lim_{s \to -\infty} X^s(t) = \infty$ *a.s.*

(2) *With* $Q(t) = X^0(t)$, $\lim_{t \to \infty} P(Q(t) < i) = 0$ *for every* $i \in \mathscr{Z}^+$.

(3) $Q(t)$ *visits state i infinitely often* (*i.o.*) *almost surely.*

Proof. (1) The event $\{X(t) = \infty\}$ is invariant with respect to the transformation Θ_s. Thus it has probability 0 or 1. Suppose $P(X(t) < \infty) = 1$. Then, by Fact 1, $X(t)$ is a stationary and ergodic process. Since $\mathbf{A}$ is a process of rate 1, $E(N_a(0, \delta]) = \delta$ for any $\delta \in \mathscr{R}$. Moreover, if $0 < \delta < 1$, then clearly this implies that $P(N_a(0, \delta] = 0) > 0$. Taking $\delta = 0.5$, we therefore get that $P(N_a(0, 0.5] = 0) > 0$. Let $B_i = \{X(0) = i;\ N_a(0, 0.5] = 0\}$. Then $P(B_i) > 0$ for some i. The event $B_i \cap \{N_s(0, 0.5] > i\}$ is contained in the event $\{X(0.5) = 0\}$ and so

$$\begin{aligned} P(X(0.5) = 0) &\geqslant P(B_i \cap \{N_s(0, 0.5] > i\}) \\ &= P(B_i) P(N_s(0, 0.5] > i) \\ &> P(B_i) \mathrm{e}^{-0.5} \frac{(0.5)^i}{i!} > 0, \end{aligned}$$

where the second equality follows from the independence of $\mathbf{S}$, for $t > 0$, with $X(0)$ and $\mathbf{A}$. And because the process $X(t)$ is stationary, $P(X(t) = 0) > 0$ for all t. In particular, $P(X(0) = 0) > 0$. Thus, $P(X(t) < \infty) = 1$ implies $P(X(0) = 0) > 0$. It now follows from [1, Lemma 2.3.1, p. 78] that this is impossible since for a critically loaded, queue, $P(X(0) = 0) = 0$ and (1) is proved.

(2) Next, since $Q(t) = X^0(t) \overset{\mathscr{D}}{=} X^{-t}(0)$, $P(Q(t) < i) = P(X^{-t}(0) < i)$. And $X^{-t}(0) \nearrow X(0) = \infty$ as $t \to \infty$. Therefore, $\lim_{t \to \infty} P(Q(t) < i) \searrow 0$ for every $i \in \mathscr{Z}^+$.

(3) Since $Q(t) \xrightarrow{\mathscr{D}} \infty$, it is enough to show that $P(Q(t)$ visits 0 i.o.$) = 1$, for then this implies $P(Q(t)$ visits i i.o.$) = 1$ for each $i \in \mathscr{Z}^+$. The event $\{Q(t) = 0$ i.o.$\}$ is Θ_s-invariant $\Rightarrow P(Q(t) = 0$ i.o.$) = 0$ or 1. If $P(Q(t) = 0$ i.o.$) = 0$, then $Q(t) > 0$ eventually with probability 1. Let $T < \infty$ be the last time that $Q(\cdot)$ transits from state 0 to state 1, never to return to 0 again. Then obviously T must coincide with an arrival of $\mathbf{A}$, and we emphasize this by setting $T = T_A$. Since $N_a(T_A, t] \geqslant N_s(T_A, t] + 1 \ \forall t > T_A$ almost surely (because $Q(t) > 0 \ \forall t > T_A$), it must be the case that the point T_A belongs to the process $\mathbf{B}$ where $\mathbf{B}$ is as defined in Lemma 1. Therefore, the process $\mathbf{B}$ has at least one point per sample path *almost surely*, i.e. the event $\{\omega : \mathbf{B}(\omega)(\mathscr{R}) > 0\}$ has probability 1. This contradicts the conclusion of Lemma 1 and Lemma 2 is proved. □

We are now ready to show that the output of a critical G/M/1 queue converges weakly to a rate 1 Poisson process. Specifically, we start processing arrivals at time 0, the arrival buffer having been empty before then. Suppose that the departure process is $\mathbf{D}^0 = \sum_{n=1}^{\infty} \delta_{t_n^{\mathrm{d}}}$ and look at $\mathbf{D}_t^0 = \Theta_t\, (\mathbf{D}^0 1_{\{t_n^{\mathrm{d}} \geqslant t\}})$. $\mathbf{D}_t^0$ is $\mathbf{D}^0$ viewed from time t onwards and to obtain it from $\mathbf{D}^0$, we first ignore all points of $\mathbf{D}^0$ that are less than t and then shift $\mathbf{D}^0$ by t units to the left. We will show that $\mathbf{D}_t^0$ goes weakly to a Poisson process as $t \to \infty$. In a sense, this is expected because Lemma

2 shows that eventually the chance that the arrival buffer is empty is going to zero and whenever the arrival buffer is non-empty, the departure process is simply the service process **S**, and this is Poisson.

We first recall the definition and some facts concerning the stochastic intensity of point processes which are of relevance to us in the sequel. Our main reference is [2].

Definition. Let **Z** be a point process adapted to some history $\mathscr{G}_t$, and let λ_t be a non-negative $\mathscr{G}_t$-progressive process such that for all $t \geqslant 0$, $\int_0^t \lambda_s \, \mathrm{d}s < \infty$ a.s. If for all non-negative $\mathscr{G}_t$-predictable processes C_t, the equality

$$E\left[\int_0^\infty C_s \, \mathrm{d}\mathbf{Z}\right] = E\left[\int_0^\infty C_s \lambda_s \, \mathrm{d}s\right]$$

holds, then **Z** admits the $\mathscr{G}_t$-intensity λ_t.

Fact 3 (Chapter II, Brémaud [2]). *Let* **Z** *have* $\mathscr{F}_t$-*intensity* λ_t *and let* $\mathscr{G}_t$ *be some history such that* $\mathscr{G}_\infty$ *is independent of* $\mathscr{F}_t$ *for all* $t \geqslant 0$. *Then* λ_t *is also the* $\mathscr{F}_t \vee \mathscr{G}_t$-*intensity of* **Z**, *where* $\mathscr{F}_t \vee \mathscr{G}_t$ *is the smallest* σ-*algebra containing both* $\mathscr{F}_t$ *and* $\mathscr{G}_t$.

Now, for a G/M/1 queue, the stochastic intensity of the service process **S** with respect to its own history $\mathscr{F}_t^S = \sigma(t_n^{\mathrm{s}}; 0 \leqslant t_n^{\mathrm{s}} \leqslant t)$ is constant and is equal to 1 (because **S** is a Poisson process). But, by Fact 3 and the independence of **S** with **A** for all $t \geqslant 0$, the intensity of **S** with respect to the combined history $\mathscr{F}_t^{A,S} = \sigma(t_n^{\mathrm{a}}, t_n^{\mathrm{s}}; 0 \leqslant t_n^{\mathrm{s}} \leqslant t, 0 \leqslant t_n^{\mathrm{a}} \leqslant t)$ is also 1. Since the process $1_{\{Q(s^-)=0\}}$ is $\mathscr{F}_t^{A,S}$-predictable (being left continuous and $\mathscr{F}_t^{A,S}$-adapted, this implies that

$$E\left[\int_0^\infty 1_{\{Q(s^-)=0\}} \, \mathrm{d}\mathbf{S}\right] = E\left[\int_0^\infty 1_{\{Q(s^-)=0\}} \, \mathrm{d}s\right]. \tag{5}$$

We now recall the definition of weak convergence of point processes [4, Chapter 9]. The sequence of point processes $\{\mathbf{D}_t^0\}$ *converges weakly* to the point process **P** as $t \to \infty$ iff for every bounded continuous function f with compact support, the random variable

$$\int_0^\infty f(s) \, \mathrm{d}\mathbf{D}_t^0 \quad \text{converges in distribution to}$$

$$\int_0^\infty f(s) \, \mathrm{d}\mathbf{P} \quad \text{as } t \to \infty.$$

Theorem 1. *The departure process of a critical G/M/1 queue,* $\mathbf{D}_t^0$, *converges weakly to a Poisson process as* $t \to \infty$.

Proof. Let f be bounded and continuous with support in $[0,N]$. By a change of variable rewrite $x_t = \int_0^\infty f(s) \, \mathrm{d}\mathbf{D}_t^0$ as $\int_t^\infty f(s-t) \, \mathrm{d}\mathbf{D}^0$. Now consider the term $y = \int_0^\infty f(s) \, \mathrm{d}\mathbf{S}$. Since the process **S** is Poisson, we are done if we can show that x_t converges in distribution to y. Since $\int_0^\infty f(s) \, \mathrm{d}\mathbf{S} = \int_t^\infty f(s-t) \, \mathrm{d}\mathbf{S} = y_t$, say, this is equivalent to showing that, as $t \to \infty$, $y_t - x_t$ goes to zero in distribution.

Now

$$\begin{aligned} y_t - x_t &= \int_t^\infty f(s-t) \, \mathrm{d}(\mathbf{S} - \mathbf{D}^0) \\ &= \int_t^{t+N} f(s-t) \, \mathrm{d}(\mathbf{S} 1_{\{Q(s^-)=0\}}), \end{aligned}$$

since $\mathbf{D}^0 = \mathbf{S} 1_{\{Q(s^-)=0\}}$ (this follows from Eq. (4)) and the support of f is in $[0,N]$. Therefore,

$$\begin{aligned} E|y_t - x_t| &\leqslant E\left[\int_t^{t+N} |f(s-t)| \, \mathrm{d}(\mathbf{S} 1_{\{Q(s^-)=0\}})\right] \\ &\leqslant |f|_{\max} E\left[\int_t^{t+N} \mathrm{d}(\mathbf{S} 1_{\{Q(s^-)=0\}})\right] \\ &= |f|_{\max} E\left[\int_t^{t+N} 1_{\{Q(s^-)=0\}} \, \mathrm{d}\mathbf{S}\right] \\ &= |f|_{\max} 1 E\left[\int_t^{t+N} 1_{\{Q(s^-)=0\}} \, \mathrm{d}s\right] \\ &= |f|_{\max} \int_t^{t+N} P(Q(s^-) = 0) \, \mathrm{d}s, \end{aligned} \tag{6}$$

where the second last equality follows from Eq. (5). The last equality follows from Fubini's theorem. All other relations are obvious. By Lemma 2,

$$\lim_{s \to \infty} P(Q(s^-) = 0) = 0.$$

Hence, by dominated convergence the right most term in (6) goes to zero as $t \to \infty$. This implies that $y_t - x_t \to 0$ in probability and hence in distribution, as was required. □

From the proof, it is apparent that the advantage in viewing the departure process of a ·/M/1 node as given by equation (4) is that we have a *bona fide* Poisson process to which the departure process converges weakly, namely the service process **S**. Since weak convergence of point processes is a distributional-type convergence, one wonders whether the stronger mode of convergence (finite time coupling) is possible. Indeed, as already mentioned, if the ·/M/1 node were stable this does happen. In Theorem 2 we use the Central Limit Theorem and show that for critical GI/M/1 queues with i.i.d. inter-arrival times having mean 1 and variance σ_{a}^2, the departure process does not couple in finite time with a Poisson process although it converges weakly to one.

Before we proceed to Theorem 2, we state the definition of finite time coupling for point processes as it applies to our problem (see also [1, Section 4.2]). As before, let $\mathbf{D}^0$ be given by $\mathbf{D}^0 = \sum_{n=1}^{\infty} \delta_{t_n^{\mathrm{d}}}$. Also suppose that $\mathbf{Z} = \sum_{n=1}^{\infty} \delta_{t_n^{\mathrm{z}}}$ is a rate 1 Poisson process. Then $\mathbf{D}^0$ is said to *couple in finite time with* $\mathbf{Z}$ if there is a random $N \in \mathscr{Z}^+$ such that $t_n^{\mathrm{d}} = t_n^{\mathrm{z}}$ for all $n > N$. That is, the departures coincide with the occurrences of a rate 1 Poisson process eventually. We now proceed to show that such a coupling is impossible.

Theorem 2. *Consider a critical GI/M/1 queue with i.i.d. mean 1, variance σ_{a}^2, inter-arrival times. The departure process from this queue does not couple in finite time with a rate 1 Poisson process. Hence for such critical GI/M/1 queues, only weak convergence of departures is possible.*

Proof. We will argue by contradiction. Suppose $\mathbf{D}^0$ couples with a (rate 1) Poisson process $\mathbf{Z}$, i.e. $\exists N(\omega) < \infty$ such that $t_n^{\mathrm{d}} = t_n^{\mathrm{z}}$ for all $n > N$. Since the inter-occurrence times of $\mathbf{Z}$ are i.i.d. exponentials with parameter 1, t_n^{z} is the sum of n i.i.d. exponentials. By the Central Limit Theorem we then get that

$$\frac{t_n^{\mathrm{z}} - n}{\sqrt{n}} \xrightarrow{\mathscr{D}} \mathscr{N}(0,1).$$

But because of the supposed coupling of $\mathbf{D}^0$ with $\mathbf{Z}$, $(t_n^{\mathrm{d}} - n)/\sqrt{n} = (t_n^{\mathrm{z}} - n)/\sqrt{n}$ for $n > N$ and so

$$\frac{t_n^{\mathrm{d}} - n}{\sqrt{n}} \xrightarrow{\mathscr{D}} \mathscr{N}(0,1) \Rightarrow P(t_n^{\mathrm{d}} \leqslant n + \varepsilon\sqrt{n}) \to \Phi(\varepsilon), \quad (7)$$

where $\Phi(\varepsilon) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\varepsilon} \mathrm{e}^{-x^2/2}\,\mathrm{d}x$.

On the other hand, given $T \in \mathscr{R}^+$, the fact that there have been n departures in $(0, T]$ implies that there have been at least n arrivals and n services in $(0, T]$. In other words, for every T

$$\{t_n^{\mathrm{d}} \leqslant T\} \subset \{t_n^{\mathrm{a}} \leqslant T\} \cap \{t_n^{\mathrm{s}} \leqslant T\}.$$

Taking $T = n + \varepsilon\sqrt{n}$ in the above equation and using the fact that the arrival and service processes are independent, we get that

$$P(t_n^{\mathrm{d}} \leqslant n + \varepsilon\sqrt{n}) \leqslant P(t_n^{\mathrm{a}} \leqslant n + \varepsilon\sqrt{n}) P(t_n^{\mathrm{s}} \leqslant n + \varepsilon\sqrt{n}). \quad (8)$$

Now $t_n^{\mathrm{a}} = t_1^{\mathrm{a}} + \sum_{i=2}^{n} t_i^{\mathrm{a}} - t_{i-1}^{\mathrm{a}}$. By the i.i.d. assumption of the inter-arrival times, t_n^{a} is a delayed renewal process (because t_1^{a} is not necessarily a mean 1, variance σ_{a}^2 random variable). Thus, by the Central Limit Theorem,

$$P(t_n^{\mathrm{a}} \leqslant n + \varepsilon\sqrt{n}) = P\left(\frac{t_n^{\mathrm{a}} - n}{\sigma_{\mathrm{a}}\sqrt{n}} \leqslant \frac{\varepsilon}{\sigma_{\mathrm{a}}}\right) \to \Phi\left(\frac{\varepsilon}{\sigma_{\mathrm{a}}}\right).$$

Using this in (8) and remembering that t_n^{s} is the sum of n i.i.d., mean 1, exponentials, we get

$$\limsup_{n\to\infty} P(t_n^{\mathrm{d}} \leqslant n + \varepsilon\sqrt{n}) \leqslant \Phi\left(\frac{\varepsilon}{\sigma_{\mathrm{a}}}\right) \Phi(\varepsilon) < \Phi(\varepsilon).$$

This contradicts (7) and hence no such coupling exists. □

3. Conclusions and final remarks

In this letter, we have identified and studied a singular feature exhibited by queues in criticality. A shift is observed in the convergence mode of the departure process from strong (pathwise coupling) to weak at the point of critical loading. Although Theorem 2 has been shown to hold under renewal arrival processes, we believe that it is true under general stationary and ergodic arrivals. However, some technical difficulties need to be resolved before such a result can be obtained.

References

[1] F. Baccelli and P. Brémaud, *Elements of Queueing Theory* (Springer, Berlin, 1994).

[2] P. Brémaud, *Point Processes and Queues: Martingale Dynamics*, Springer Series in Statistics (Springer, New York, 1981).

[3] P.J. Burke, The output of a queueing system, *Oper. Res.* **4** (1956) 699–704.

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

[5] F.P. Kelly, *Reversibility and Stochastic Networks* (Wiley, London, 1979).

[6] L. Kleinrock, *Queueing Systems*, Vols. 1 and 2 (Wiley, New York, 1964).

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

[8] B. Prabhakar, N. Bambos and T.S. Mountford, On the synchronization of Poisson processes and queueing networks with service and synchronization nodes, filed as UCLA Technical Report: Eng-94-001 (1994).

[9] E. Reich, Waiting times when queues are in tandem, *Ann. Math. Statist.* **28** (1957) 768–773.