Queueing Syst (2011) 67: 111–126
DOI 10.1007/s11134-010-9203-0

# Large number of queues in tandem: Scaling properties under back-pressure algorithm

**Alexander L. Stolyar**

Received: 30 October 2009 / Revised: 9 November 2010 / Published online: 30 November 2010


**Abstract** We consider a system with $N$ unit-service-rate queues in tandem, with exogenous arrivals of rate $\lambda$ at queue 1, under a back-pressure (MaxWeight) algorithm: service at queue $n$ is blocked unless its queue length is greater than that of the next queue $n+1$. The question addressed is how steady-state queues scale as $N \to \infty$. We show that the answer depends on whether $\lambda$ is below or above the critical value $1/4$: in the former case the queues remain uniformly stochastically bounded, while otherwise they grow to infinity.

The problem is essentially reduced to the behavior of the system with an infinite number of queues in tandem, which is studied using tools from interacting particle systems theory. In particular, the criticality of load $1/4$ is closely related to the fact that this is the maximum possible flux (flow rate) of a stationary totally asymmetric simple exclusion process.

**Keywords** Queueing networks · Interacting particle systems · Stability · Back-pressure · MaxWeight · Infinite tandem queues · TASEP

**Mathematics Subject Classification (2000)** 90B15 · 60K25 · 60K35 · 68M12

## 1 Introduction

In this paper, we consider a system with $N$ single-server queues in tandem, with exogenous Poisson arrival process at queue 1 and customers leaving after service in queue $N$. All service times have exponential distribution with unit mean. The waiting room in each queue is unlimited, but the system operates under the following back-pressure policy: service of queue $n$ is blocked (stopped) unless its queue length is

A.L. Stolyar (✉)
Bell Labs, Alcatel-Lucent, 600 Mountain Ave., 2C-322, Murray Hill, NJ 07974, USA
e-mail: stolyar@research.bell-labs.com

greater than that of the "next-hop" queue $n+1$. The system is stable as long as the input rate $\lambda < 1$. The main question we address is: Given that the load $\lambda < 1$ is fixed, how do the steady-state queue lengths increase (scale) as $N \to \infty$.

Our main motivation comes from the fact that general *back-pressure* (BP) policies (sometimes also called *MaxWeight*), originally introduced by Tassiulas and Ephremides [13] and having received much attention in the literature (cf. [4, 12] for recent reviews), are very attractive for application in communication and service networks. This is due to the adaptive nature of BP policies—they can ensure maximum possible network throughput, without a priori knowledge of flow input rates. A key mechanism in BP policies, giving them the adaptivity (and the name), is that the "priority" of a traffic flow $f$ at a given network node $n$ is "proportional" to the difference between flow $f$ queue lengths at node $n$ and the next node on the route; in particular, unless this queue-differential is positive, flow $f$ service at $n$ can be blocked. This mechanism, however, has a drawback, namely it can lead to a large queue build-up along a flow route, since, roughly speaking, the queue "needs" to increase as we move back from a flow destination node to its source. Such "bad scaling" behavior of BP algorithms is emphasized in [3], and is not very surprising (see also our Proposition 2 and bound (3)). One approach to mitigate this scaling problem in practical systems is proposed in [2, 3], where it is suggested to "run" BP algorithm on virtual queues as opposed to physical ones.

It is of interest to understand fundamental scaling properties of BP algorithms. In this paper we address this problem for a simple system—a single flow served by $N$ queues in tandem. We show that, perhaps somewhat surprisingly, the scaling behavior of BP is *not always* bad. Namely, if load is below some critical level, $\lambda < 1/4$ for our model, all queues remain uniformly stochastically bounded for all $N$. (In fact, we show that the stochastic bound has an exponentially decaying tail; see Theorem 1(i).) When $\lambda > 1/4$, the queues increase to infinity with $N$ (see Theorem 1(ii)).

The problem of asymptotic behavior of queues as $N \to \infty$ is essentially reduced (see Proposition 5) to the behavior of the system with an *infinite* number of queues in tandem. Such an infinite-tandem model is within the framework of interacting particle systems [8, 9]—methods and results of the corresponding theory will be our main tools. As we will see, the criticality of load $1/4$ is closely related to the fact that $1/4$ is the maximum possible flux (average flow rate) of a stationary *totally asymmetric simple exclusion process* (TASEP). In the subcritical case, $\lambda < 1/4$, a stationary regime exists such that only a finite (random) number of "left-most" queues can be greater than 1; the rest of the queues have at most one customer and the process "there" behaves like TASEP (see Theorems 10 and 11). In the supercritical case, $\lambda > 1/4$, each queue grows without bound with time (see Theorem 12).

Infinite series of queues in tandem is a much studied model in the literature (cf. [11] and references therein), under a variety of assumptions. In particular, [11] studies an infinite system under a class of blocking policies (which are different from the BP policy in our model), where blocking is caused by limited waiting space (finite buffer) in the queues; there, the phenomenon of critical load, below which the system is stable, also exists.

The rest of the paper is organized as follows. Section 2 presents the formal model and main result. The "reduction" of our problem to the behavior of the infinite system,

and basic properties of the latter, are given in Sect. 3. The subcritical and supercritical load cases are treated in Sects. 4 and 5, respectively. In Sect. 6 we remark on more general input processes. We conclude in Sect. 7. The appendix contains some of the proofs.

## 2 Formal model and main result

Consider a series of $N$ servers (sites), numbered $1, \ldots, N$, each with unlimited queueing room. A Poisson flow of customers, of rate $\lambda > 0$, arrives at server 1, and each customer has to be served consecutively by the series of servers, from 1 to $N$; after service by the $N$th server, a customer leaves the system. The service time of any customer at each server is exponentially distributed, with mean value 1; all service times are independent of each other and of the input process. To be specific, assume that each site serves customers in first-come-first-serve order—given Markov assumptions, this will not limit the generality of results.

Let us denote by $Q_n^{(N)}(t), n = 1, \ldots, N, t \geq 0$, the queue length at the $n$th server at time $t$. The superscript $N$ indicates the number of servers, which will be the parameter we vary.

Consider the following *back-pressure* (BP) algorithm: site $n$ is actually serving a customer (which is the head-of-the-line customer from its queue) at time $t$ if and only if $Q_n^{(N)}(t) > Q_{n+1}^{(N)}(t)$. (We use the convention that $Q_{N+1}^{(N)}(t) \equiv 0$.) In other words, the service at site $n$ is blocked unless the queue at site $n+1$ is smaller.

The random process $(Q^{(N)}(t) \equiv (Q_n^{(N)}(t), n = 1, 2, \ldots, N), t \geq 0)$, describing evolution of the queues, is a countable irreducible, continuous-time Markov chain. Stability of this process—ergodicity of the Markov chain—is guaranteed under the condition $\lambda < 1$—this follows from well-known properties of BP algorithms (cf. [4]). Therefore, the unique stationary distribution exists; we denote by $Q^{(N)}(\infty)$ a random system state in the stationary regime.

The question we address is how the steady-state queues $Q_n^{(N)}(\infty)$ grow (scale) as $N \to \infty$; more specifically whether or not they remain stochastically bounded. Our main result is the following

**Theorem 1**

(i) *If* $\lambda < 1/4$, *then there exist* $C_1 > 0$ *and* $C_2 > 0$ *such that, uniformly in* $N$,

$$\mathbb{P}\left\{ \max_{1 \leq n \leq N} Q_n^{(N)}(\infty) > r \right\} \leq C_1 e^{-C_2 r}. \tag{1}$$

(ii) *If* $\lambda > 1/4$, *then for any* $n \geq 1$, $Q_n^{(N)}(\infty) \to \infty$ *in probability as* $N \to \infty$.

Statements (i) and (ii) will follow from Theorems 10 and 12, respectively, which are concerned with the corresponding infinite-tandem system, defined in Sect. 3. (It will follow from Propositions 5 and 6 that it suffices to concentrate on the behavior of the first queue of the infinite system.)

Note that the fact that the tight uniform bound (1) *cannot* possibly hold for all $\lambda < 1$ is fairly obvious. First, it is easy to observe that if the condition

$$Q_n^{(N)}(t) + 1 \geq \max_{k>n} Q_k^{(N)}(t), \quad \forall n \geq 1, \tag{2}$$

holds for $t = t_0$, then it holds for all $t \geq t_0$ as well; in particular, it does hold in the stationary regime. Secondly, in the stationary regime the average rate at which customers move from site $n$ to $n+1$, for any $n \leq N$, is $\lambda$. Therefore, we have

**Proposition 2** *If* $\lambda < 1$, *then for any* $n = 1, \ldots, N$,

$$\mathbb{P}\big\{Q_n^{(N)}(\infty) > Q_{n+1}^{(N)}(\infty)\big\} = \lambda.$$

*Then*, *using* (2),

$$\mathbb{E}Q_n^{(N)}(\infty) - \mathbb{E}Q_{n+1}^{(N)}(\infty) \geq 2\lambda - 1, \quad n = 1, \ldots, N.$$

Thus, in the case $\lambda > 1/2$, we have at least linear growth of the first queue expected value:

$$\mathbb{E}Q_1^{(N)}(\infty) \geq (2\lambda - 1)N. \tag{3}$$

## 3 Basic properties of infinite-tandem queues under the back-pressure algorithm

Consider a system just like the one in Sect. 2, except it has an infinite number of servers (sites) in tandem, indexed $n = 1, 2, \ldots$. Arriving customers never leave–they just keep moving from site to site, to the "right". We denote by $Q_n(t)$, $n = 1, 2, \ldots$, $t \geq 0$, the queue length at site $n$ at time $t$, by $Q(t) \equiv (Q_n(t), n = 1, 2, \ldots)$ the state of the entire system at time $t$. It will be convenient to assume that the phase space for each queue (site) $n$ state $Q_n$ is the compact set $\bar{\mathbb{Z}}_+ = \mathbb{Z}_+ \cup \{\infty\}$ with $\mathbb{Z}_+$ being the set of non-negative integers and with metric, for instance, $|e^{-i} - e^{-j}|$. (To make the back-pressure algorithm well-defined when some queues may be infinite, we use the following conventions: an infinite queue remains infinite when a customer leaves it after a service completion or a new customer arrives at it—this implies that an infinite queue stays infinite at all times; if $Q_n(t) = Q_{n+1}(t) = \infty$ the service at site $n$ is blocked.) The state space of Markov process $(Q(t), t \geq 0)$ is $\bar{\mathbb{Z}}_+^S$, $S = \{1, 2, \ldots\}$, with product topology; the process is formally defined within the framework of interacting particle systems (cf. Sect. I.3 of [8], specifically Theorem I.3.9). We will use some slightly abusive notations: $Q(\cdot)$ for the process $(Q(t), t \geq 0)$, and $Q = (Q_n, n = 1, 2, \ldots) \in \bar{\mathbb{Z}}_+^S$ for elements of the phase space; and will denote $\|Q\| = \sum_n Q_n$.

Throughout the paper we will also use the following constructive representation of process $Q(\cdot)$, which is standard for Markov interacting-particle systems and is convenient for coupling (cf. [8, 9]) of different versions of the processes. The underlying probability space is such that there is a unit rate Poisson processes $\Pi_n$ for each site $n$; these processes are independent from each other and of the input flow Poisson

process. Then, if time $\tau \geq 0$ is a point of the Poisson process $\Pi_n$, a customer moves from queue $n$ to queue $n+1$ at $\tau$ if $Q_n(\tau-) > Q_{n+1}(\tau-)$, otherwise the move is suppressed. The sample paths of the process $Q(\cdot)$, constructed this way, are well defined with probability 1. This fact is immediate when $\|Q(0)\| < \infty$. If $\|Q(0)\| = \infty$, it suffices to observe that w.p.1 for any $t \geq 0$ and any $n \geq 1$ there exists $n' > n$ such that there are no points of process $\Pi_{n'}$ in $[0,t]$; therefore, w.p.1 the realization of $(Q_1(\tau), \ldots, Q_n(\tau))$ in the interval $[0,t]$ is uniquely defined by the finite number of points of processes $\Pi_1, \ldots, \Pi_{n'-1}$ and the exogenous arrival process in this interval.

The finite system of Sect. 2, with $N$ sites, will be viewed as an infinite one, but with the modification that any customer reaching site $N+1$ is immediately removed from the system, and with $Q_n(t) \equiv 0$ for $n > N$.

**Lemma 3** *An analog of* (2) *holds for the infinite system. Namely, if the condition*

$$Q_n(t) + 1 \geq \sup_{k>n} Q_k(t), \quad \forall n \geq 1, \tag{4}$$

*holds for* $t = 0$, *then w.p.*1 *it holds for all* $t \geq 0$.

Proof is in Appendix A.

Now we state basic monotonicity properties of the infinite system. Additional notation: the inequality $Q \leq Q'$ is understood component-wise; the partial order relation $Q \preceq Q'$ means that both $\|Q\|$ and $\|Q'\|$ are finite and

$$\sum_{k \geq n} Q_k \leq \sum_{k \geq n} Q'_k, \quad \forall n \geq 1;$$

if $Q$ and $Q'$ are random elements in the phase space $\bar{\mathbb{Z}}_+^S$, $Q \leq_{\mathrm{st}} Q'$ means that $Q$ is stochastically dominated by $Q'$ in the sense of $\leq$ order, i.e., $Q$ and $Q'$ can be constructed on a common probability space in a way such that $Q \leq Q'$ w.p.1; the symbol $\Rightarrow$ denotes convergence in distribution of random elements (in a space that will be clear from the context), i.e., the weak convergence of their distributions.

**Lemma 4** *Process* $Q(\cdot)$ *is* monotone [8] *with respect to both* $\leq$ *and* $\preceq$ *partial order*; *namely, if* $Q(0) \leq Q'(0)$ [*respectively*, $Q(0) \preceq Q'(0)$], *then the processes* $Q(\cdot)$ *and* $Q'(\cdot)$ *can be coupled so that* $Q(t) \leq Q'(t)$ [*respectively*, $Q(t) \preceq Q'(t)$] *for all* $t \geq 0$. *The same is true for the process* $Q^{(N)}(\cdot)$ *with any* $N$.

Proof is in Appendix B. Essentially as a corollary of Lemmas 4 and 3, we obtain the following

**Proposition 5** *Consider the infinite system and the finite systems, for each* $N = 1, 2, \ldots$, *all with zero initial state* (*with all queues being* 0). *Then*,

(i) *All corresponding processes can be coupled so that for all* $t \geq 0$,

$$Q^{(1)}(t) \leq Q^{(2)}(t) \leq \cdots \leq Q^{(N)}(t) \leq \cdots \leq Q(t), \tag{5}$$

*and for each* $t \geq 0$, $Q^{(N)}(t) \uparrow Q(t)$ *as* $N \to \infty$.

(ii) *Process $Q(\cdot)$ is stochastically non-decreasing in $t$ (in the sense of $\leq$ order), and process $Q^{(N)}(\cdot)$ is stochastically non-decreasing in both $t$ and $N$; in other words, for any $t_1 \leq t_2$ and $N_1 \leq N_2$, $Q(t_1) \leq_{\mathrm{st}} Q(t_2)$ and $Q^{(N_1)}(t_1) \leq_{\mathrm{st}} Q^{(N_2)}(t_2)$.*

(iii) *We have*

$$Q^{(N)}(t) \Rightarrow Q^{(N)}(\infty), \quad \text{as } t \to \infty, \tag{6}$$

*where the distribution of $Q^{(N)}(\infty)$ is the stationary distribution of process $Q^{(N)}(\cdot)$, and $Q^{(N)}(\infty)$ is stochastically non-decreasing in $N$;*

$$Q^{(N)}(\infty) \Rightarrow Q(\infty), \quad \text{as } N \to \infty, \tag{7}$$

*where $Q(\infty)$ is a random element in $\bar{\mathbb{Z}}_+^S$;*

$$Q(t) \Rightarrow Q(\infty), \quad \text{as } t \to \infty. \tag{8}$$

(iv) *The distribution of $Q(\infty)$ is a stationary distribution, moreover—the lower invariant measure, of Markov process $Q(\cdot)$.*

(v) *Condition (4) holds for all $t \geq 0$ and for the stationary state $Q(\infty)$.*

(vi) *For each finite $t \geq 0$ and $t = \infty$, define*

$$B\big(Q(t)\big) \doteq \min\big\{n = 1, 2 \ldots \mid Q_n(t) = 0\big\}. \tag{9}$$

*Then, random variable $B(Q(t))$ (with values in $\bar{\mathbb{Z}}_+$) is stochastically non-decreasing in $t \geq 0$, and $B(Q(t)) \Rightarrow B(Q(\infty))$ as $t \to \infty$.*

Proof is in Appendix C.

**Proposition 6**

(i) *Consider process $Q(\cdot)$ with a fixed initial state $Q(0)$ such that, for some finite $n \geq 2$, $Q_{n-1}(0) = \infty$ and $Q_n(0) < \infty$. Then $Q_n(t) \to \infty$ in probability as $t \to \infty$.*

(ii) *Consider the stationary (random) state $Q(\infty)$ of the process $Q(\cdot)$. If $\mathbb{P}\{Q_1(\infty) = \infty\} > 0$ or $\mathbb{P}\{B(Q(\infty)) = \infty\} > 0$, then $Q_n(\infty) = \infty$ w.p.1 for all $n \geq 1$.*

Proof is in Appendix D.

Propositions 5 and 6 show, in particular, that to prove Theorem 1, we can study the distribution of $Q_1(\infty)$. Indeed, $Q_1(\infty)$ is the limit (in distribution) and stochastic upper bound of $Q_1^{(N)}(\infty)$. If $Q_1(\infty) < \infty$ w.p.1, $Q_1(\infty) + 1$ is a uniform stochastic upper bound on each $Q_n^{(N)}(\infty)$; if $Q_1(\infty) = \infty$ with non-zero probability, then for each $n$, $Q_n^{(N)}(\infty) \to \infty$ in probability as $N \to \infty$.

We will need one more monotonicity property, which is also a corollary of Lemma 4. Its meaning is very simple: if in addition to process $Q(\cdot)$ we consider another process $Q'(\cdot)$, which is constructed the same way as $Q(\cdot)$, but with some additional exclusions ("obstructions") on the movement of the customers, then $Q'(\cdot)$ will stay "behind" $Q(\cdot)$ in the sense of $\preceq$ order.

**Proposition 7** *Consider a fixed realization of the process $Q(\cdot)$, determined by a fixed initial state $Q(0)$, $\|Q(0)\| < \infty$, and realizations of the exogenous arrival process (at site 1) and of all processes $\Pi_n$, $n \geq 1$. Assume the realization is well defined in that there is a finite number of transitions in any finite interval. Suppose further that in the realizations of processes $\Pi_n$, some of the points (jumps) are marked as "valid" (in an arbitrary way) and the remaining points are "invalid". Consider another realization $Q'(\cdot)$, with the same initial state $Q'(0) = Q(0)$, and constructed in the same way as $Q(t)$, except the invalid points of processes $\Pi_n$ are "ignored" (cause no action). Then,*

$$Q'(t) \preceq Q(t), \quad \forall t \geq 0,$$

*and, in particular, $Q'_1(t) \geq Q_1(t)$ for all $t$.*

Proof is in Appendix E.

## 4 Subcritical case: $\lambda < 1/4$

Suppose $\lambda < 1/4$. We will construct a process $Q'(\cdot)$, coupled with $Q(\cdot)$ so that Proposition 7 holds path-wise, and such that we can obtain a stochastic upper bound on $Q'_1(t)$.

Consider process $Q(\cdot)$, with zero initial state $\|Q(0)\| = 0$, constructed on the probability space described in Sect. 3. We will extend the probability space to define a stationary *totally asymmetric simple exclusion process* (TASEP, cf. Chap. VIII of [8]), with sites being integers $n \in \mathbb{Z}$, and particles moving to the "right". Specifically, assume that there is a site associated with each integer $n \in \mathbb{Z}$, not just positive $n$, and augment the probability space so that an independent, unit rate Poisson processes $\Pi_n$ is associated with each site $n \in \mathbb{Z}$. (Processes $\Pi_n$ with $n \leq 0$ do *not* affect process $Q(\cdot)$.) Let us choose arbitrary (density) $\rho \leq 1/2$, such that $\mu = \rho(1-\rho) > \lambda$. Let $Y_n(t) \in \{0, 1\}$ denote the number of particles of TASEP at site $n \in \mathbb{Z}$ at time $t \geq 0$. We further augment the probability space so that, independently of all other driving processes, at time 0 each site $n \in \mathbb{Z}$ contains a particle, $Y_n(0) = 1$, with probability $\rho$ and does not contain one, $Y_n(0) = 0$, with probability $1 - \rho$. The movement of TASEP particles will be driven by the Poisson processes $\Pi_n$. (Note that $\Pi_n$ with $n \geq 1$ are the same processes that drive $Q(\cdot)$. On the other hand, the exogenous input process at site 1 does *not* affect TASEP.) If time $\tau$ is a point (jump) of $\Pi_n$ associated with site $n$, then the particle located at $n$ (if any) attempts to jump to site $n+1$ – it actually does jump if site $n+1$ is empty, and it stays at $n$ otherwise. It is well known that if the initial state $Y(0)$ has Bernoulli distribution as defined above, then the TASEP process $Y(\cdot)$ is stationary (cf. Theorem VIII.2.1 of [8]). The flux of this process, i.e., the average departure rate of particles from a given site, is $\mu = \rho(1-\rho)$; the average speed of a given ("tagged") particle is $v = 1 - \rho$. (For example, $\rho = 1/2$ gives the maximum possible flux $\mu = 1/4 > \lambda$; this is where the condition $\lambda < 1/4$ comes from: $\lambda$ needs to be less than the flux of a stationary TASEP.)

The process $Q'(\cdot)$ has the same (zero) initial state as $Q(\cdot)$, and is constructed the same way as $Q(\cdot)$ except for an additional exclusion: a customer (particle) from

queue (site) 1 cannot move to queue 2 at time $\tau$ unless a particle of the TASEP jumps from site 1 to 2 at $\tau$. We now record basic properties of process $Q'(\cdot)$.

**Proposition 8**

(i) $Q'(t) \preceq Q(t)$ *and* $Q'_1(t) \geq Q_1(t)$ *for all* $t \geq 0$.
(ii) *For any* $n \geq 2$ *and any* $t \geq 0$, $Q'_n(t) \leq Y_n(t)$. *In other words, at all sites to the right of* 1, *process* $Q'(\cdot)$ *stays "within TASEP"; in particular, there can be at most one particle in each site* $n \geq 2$.
(iii) *A particle jump from site* 1 *to* 2 *in the process* $Q'(\cdot)$ *happens at time* $\tau$ *if and only if* $Q'_1(\tau-) \geq 1$ *and there is a jump of TASEP particle from* 1 *to* 2 *at time* $\tau$.

Proof is in Appendix F.

We know from Proposition 7 that $Q'_1(t)$ is an upper bound of $Q_1(t)$. The behavior of queue length $Q'_1(t)$ is such that it is initially zero, $Q'_1(0) = 0$, the input process is Poisson with rate $\lambda$, and the "service process" is the stationary process of TASEP particle jumps from site 1 to 2. We denote by $A(t_1, t_2)$ and $S(t_1, t_2)$ the number of points (jumps) of the arrival and service processes, respectively, in the interval $(t_1, t_2]$; WLOG we assume that these processes are defined for all real times, that is $t_1, t_2 \in \mathbb{R}$, $t_1 \leq t_2$; their average rates are $\lambda$ and $\mu = ES(t_1, t_2)/(t_2 - t_1)$, respectively. From the large deviations estimate given below in Lemma 9, it also follows that $S(-s, 0)/s \to \mu$, $s \to \infty$, with probability 1.

From classical Loynes constructions [10], it is known that the distribution of $Q'_1(t)$ is stochastically non-decreasing with $t$, and as $t \to \infty$ it weakly converges to the stationary distribution, which in turn is the same as that of the random variable

$$Q'_1(\infty) \doteq \sup_{s \geq 0} \big[A(-s, 0) - S(-s, 0)\big]. \tag{10}$$

(In our case, these facts can also be seen directly, since $Q'_1(t)$ is equal in distribution to $\sup_{s \in [0,t]}[A(-s, 0) - S(-s, 0)]$.) $Q'_1(\infty)$ is a proper random variable due to condition $\lambda < \mu$, which guarantees that the RHS of (10) is finite w.p.1.

Thus, we see that, as $t \to \infty$, $Q'_1(t)$ (and then $Q_1(t)$) remains stochastically bounded by $Q'_1(\infty)$. Moreover, the large deviations estimates (in Lemma 9 below) will imply an exponential bound on the tail of the $Q'_1(\infty)$ distribution. We proceed with the details.

The following fact is a known property of the stationary TASEP defined above. At time 0, let us consider the site with smallest index $n_0 \geq 2$ that contains a particle, and tag this particle. (Obviously, $n_0 - 2$ has geometric distribution.) If we consider the point process of jumps of tagged particle in time interval $[0, \infty)$, it is a Poisson process of rate $v = 1 - \rho$ (cf. Corollary VIII.4.9 of [8]). Therefore, the location $H(t)$ of the tagged particle at time $t \geq 0$ is $H(t) = 2 + H_1 + H_2(t)$, where $H_1$ is geometric random variable with mean $(1 - \rho)/\rho$, $H_2(t)$ is Poisson r.v. with mean $vt$, and $H_1$ and $H_2(t)$ are independent. From the stationarity of TASEP we also know that, at any time $t$, the total number $G(n)$ of particles at sites $2, 3, \ldots, n$ is simply the sum of $n - 1$ independent Bernoulli variables with mean $\rho$. Using "separate" large deviations estimates for $H(t)$ and $G(n)$, even though these two r.v. are not independent, we obtain the following

**Lemma 9** *For any $\delta > 0$, there exist $C_3 > 0$ and $C_4 > 0$ such that*

$$\mathbb{P}\{|S(0,t) - \mu t| > \delta t\} \le C_3 e^{-C_4 t}. \tag{11}$$

*Proof* To prove the bound

$$\mathbb{P}\{S(0,t) < (\mu - \delta)t\} \le C_3 e^{-C_4 t}, \tag{12}$$

we can choose $\epsilon > 0$ small enough so that

$$\mathbb{P}\{S(0,t) < (\mu - \delta)t\} \le \mathbb{P}\{H(t) \le (v - \epsilon)t\} + \mathbb{P}\{G((v-\epsilon)t) \le (\rho - \epsilon)(v - \epsilon)t\}. \tag{13}$$

Indeed, if $H(t) > (v-\epsilon)t$ (which means the tagged particle is to the right of $(v-\epsilon)t$ at time $t$) and $G((v-\epsilon)t) > (\rho-\epsilon)(v-\epsilon)t$ (which means the number of particles at time $t$ between 2 and $(v-\epsilon)t$ is greater than $(\rho-\epsilon)(v-\epsilon)t$), then at least $(\rho - \epsilon)(v-\epsilon)t$ particles moved from site 1 to site 2 in time interval $[0,t]$. Since $\rho v = \mu$, for a small enough $\epsilon$, $(\rho-\epsilon)(v-\epsilon)t > (\mu - \delta)t$. This proves (13), and then (12). The bound

$$\mathbb{P}\{S(0,t) > (\mu + \delta)t\} \le C_3 e^{-C_4 t}$$

is proved similarly; namely, it follows from

$$\mathbb{P}\{S(0,t) > (\mu+\delta)t\} \le \mathbb{P}\{H(t) \ge (v+\epsilon)t\} + \mathbb{P}\{G((v+\epsilon)t) \ge (\rho+\epsilon)(v+\epsilon)t\},$$

which holds for a small $\epsilon$. □

**Theorem 10** *Assume $\lambda < 1/4$. There exist $C_1 > 0$ and $C_2 > 0$ such that*

$$\mathbb{P}\{Q_1'(\infty) > r\} \le C_1 e^{-C_2 r}, \tag{14}$$

*and then $\mathbb{P}\{Q_1(\infty) > r\} \le C_1 e^{-C_2 r}$ and $\mathbb{P}\{\sup_n Q_n(\infty) > r+1\} \le C_1 e^{-C_2 r}$.*

*Remark* Given the large deviations bound (11), the argument to prove (14) is quite standard (see [5, 6]). However, formally, [5] for example, requires a stronger condition, large deviations principle (LDP) for $S(0,t)/t$; we did not find this LDP result in the literature and it is not needed to prove (14). Therefore, for completeness, we give a proof of the theorem.

*Proof* It follows from the definition (10) that for any fixed $d > 0$

$$\begin{aligned} Q_1'(\infty) &\le \sup_{k=1,2,\ldots} \big[A(-kd,0) - S\big(-(k-1)d,0\big)\big] \\ &= A(-d,0) + \sup_{k=1,2,\ldots} \big[A(-kd,-d) - S\big(-(k-1)d,0\big)\big]. \end{aligned} \tag{15}$$

Let us fix $b > 0$ such that $b\lambda < 1$, and for each $r > 0$ we will choose $d = br$. Then we can write

$$\mathbb{P}\{Q'_1(\infty) > r\} \le \mathbb{P}\{A(-br, 0) \ge r\} + \sum_{k=2,3,\ldots} \mathbb{P}\{A(-kd, -d) - S(-(k-1)d, 0) \ge 0\}.$$

If we fix $\delta > 0$ small enough so that $\lambda + \delta < \mu - \delta$ and $(\lambda + \delta)b < 1$, we have

$$\mathbb{P}\{Q'_1(\infty) > r\} \le \mathbb{P}\{A(-br, 0) \ge (\lambda + \delta)br\} + \sum_{k=2,3,\ldots} \big[\mathbb{P}\{A(-kd, -d) \ge (\lambda + \delta)(k-1)d\} + \mathbb{P}\{S(-(k-1)d, 0) \le (\mu - \delta)(k-1)d\}\big].$$

It remains to apply large deviations bounds on $A(\cdot)$ and $S(\cdot)$ (Lemma 9). □

Let us recall that $B(Q(t))$ is defined in (9) as the index of the left-most empty site, or "busy interval", of state $Q(t)$. Note that to the right of, and including, site $B(Q(t))$ all sites have at most one customer, and so the (instantaneous) evolution of the process follows the same "rules" as that of TASEP. Let us also define

$$B'(Q(t)) \doteq \min\left\{n = 1, 2, \ldots \,\middle|\, \sum_{k=1}^{n} Q_k(t) < n\right\}.$$

The interpretation of $B'(Q(t))$ is as follows. Let us modify the state $Q(t)$ in the following way: we take a customer from the left-most site with 2 or more customers, and move it to the left-most empty site, and then repeat until all sites have at most 1 customer. If this procedure stops in a finite number of steps, $B'(Q(t))$ is the index of the left-most empty site of the modified state; otherwise, $B'(Q(t)) = \infty$.

Obviously, $B'(Q(t)) \ge B(Q(t))$; in addition, $B'(Q'(t)) \ge B'(Q(t))$ because $Q'(t) \preceq Q(t)$ and $\|Q'(t)\| = \|Q(t)\|$. Thus, $B'(Q'(t)) \ge B(Q(t))$.

**Theorem 11** *Assume $\lambda < 1/4$. There exist $C_5 > 0$ and $C_6 > 0$ such that, for all $t \ge 0$*

$$\mathbb{P}\{B'(Q'(t)) > n\} \le C_5 e^{-C_6 n},$$

*and then*

$$\mathbb{P}\{B(Q(t)) > n\} \le C_5 e^{-C_6 n}.$$

*Proof* The argument is similar to that in the proof of Lemma 9. Recall the definition of $B'(Q'(t))$ and Proposition 8(ii). They imply that for a small fixed $\epsilon > 0$

$$\{B'(Q'(t)) > n\} \subseteq \left\{\sum_{2}^{n} [1 - Y_k(t)] \le (1 - \rho - \epsilon)(n-1)\right\} \cup \{Q'_1(t) \ge (1 - \rho - 2\epsilon)(n-1)\}.$$

Now recall that $[1 - Y_k(t)]$ are i.i.d. Bernoulli with mean $1 - \rho$, and that $Q'_1(t) \leq_{\mathrm{st}} Q'_1(\infty)$. It remains to use the large deviations bounds on the $\sum_2^n [1 - Y_k(t)]$ and on $Q'_1(\infty)$ (Theorem 10). □

Theorem 11 illustrates in particular the fact that when $\lambda < 1/4$ the infinite-tandem system under BP algorithm in the stationary regime behaves in the following way: there is only a "small" number of sites (from site 1 to site $B(Q(t))$) where the queue can be greater than one, while all sites to the right of $B(Q(t))$ have queue of at most one, and therefore the behavior of the process "to the right of $B(Q(t))$" is the same as that of TASEP.

## 5 Supercritical case: $\lambda > 1/4$

Note that if $\lambda > 1/2$ we immediately see from (3) that $\mathbb{E}Q_1(\infty) = \lim_N \mathbb{E}Q_1^{(N)}(\infty) = \infty$. Here we prove that, in fact, $Q_1(\infty)$ is infinite w.p.1, under a weaker condition $\lambda > 1/4$. The intuition behind our argument is as follows. Unless $Q_1(\infty) = \infty$ w.p.1, the busy interval $B(Q(t))$ must be stochastically bounded w.p.1. Then, in the stationary regime, all sites "far enough" to the right have at most one customer (particle) in them, and therefore the process "there" behaves as TASEP, and its flux cannot exceed that of a stationary "one-sided" TASEP [7] ("living" on the positive integers). The flux of one-sided TASEP cannot be greater than $1/4$, while the flux of our process must be $\lambda > 1/4$, a contradiction.

**Theorem 12** *Assume* $\lambda > 1/4$. *Then* $Q_1(\infty) = \infty$ *w.p.*1. (*And then* $Q_n(\infty) = \infty$ *w.p.*1 *for all* $n \geq 1$.)

*Proof* The proof is by contradiction—assume $Q_1(\infty)$ is finite with positive probability. Then, by Proposition 6(ii), $Q_1(\infty) < \infty$ and $B(Q(\infty)) < \infty$ w.p.1.

The stationary version of the process $Q(\cdot)$ (i.e., the one with stationary distribution equal to that of $Q(\infty)$) we denote by $\tilde{Q}(\cdot)$. This process is such that w.p.1 condition (4) holds for all $t \geq 0$, and therefore all $\tilde{Q}_n(t)$ for all $t$ are uniformly stochastically upper bounded by $Q_1(\infty) + 1$ and then finite w.p.1; $B(\tilde{Q}(t))$ are finite w.p.1 (equally distributed) random variables for all $t$; the flux is equal to $\lambda$, namely, $\mathbb{E}F_n(t)/t = \lambda$ for any $t > 0$ and $n \geq 1$, where $F_n(t)$ is the number of customer arrivals at site $n$ in interval $(0, t]$. Note that for $n \geq 2$, the arrival process $F_n(\cdot)$ at site $n$ is the departure process from site $n - 1$. This implies that increments of processes $F_n(\cdot)$, $n \geq 2$, are stochastically upper bounded by the increments of independent Poisson processes $\Pi_{n-1}(\cdot)$.

Consider space-shifted processes $\{[T_m\tilde{Q}](\cdot), [T_m F](\cdot), [T_m \Pi](\cdot)\}$, $m = 1, 2, \ldots$, where $[T_m\tilde{Q}]_i(t) = \tilde{Q}_{m+i}(t)$, $i = 1, 2, \ldots$, $t \geq 0$, and $[T_m F]_i(t)$ and $[T_m \Pi]_i(t)$ defined similarly. For each $m$ this process is such that $[T_m\tilde{Q}](\cdot)]$ and the increments of $[T_m F](\cdot)$ and $[T_m \Pi](\cdot)$ are stationary. (Note that process $[T_m\tilde{Q}](\cdot)$ is *not* Markov.) This process is still well-defined if we assume that each component $[T_m\tilde{Q}]_i(t)$, $[T_m F]_i(t)$ and $[T_m \Pi]_i(t)$ takes values in the (non-compact) space $\mathbb{Z}_+$ with the usual

topology (because we know that they are finite w.p.1), and with corresponding product topology on the state space of the process. Note that, since $B(\tilde{Q}(t))$ is finite w.p.1 and all sites to the right of $B(\tilde{Q}(t))$ contain at most one particle, we have

$$\lim_{m\to\infty} \mathbb{P}\Big\{\sup_{n\geq 1}[T_m \tilde{Q}]_n(t) \leq 1\Big\} = 1, \quad \forall t \geq 0. \tag{16}$$

Then, using properties of process $\tilde{Q}(\cdot)$ described earlier, in particular the fact that the increments of $F_n(\cdot)$ are bounded by those of $\Pi_{n-1}(\cdot)$, it is easy to see that a process consisting of any finite subset of components $[T_m \tilde{Q}]_i(\cdot)$, $[T_m F]_i(\cdot)$ and $[T_m \Pi]_i(\cdot)$ is tight (cf. Theorem 15.2 in [1]). Consequently, there exists a subsequence of $\{m\}$ along which the shifted process converges in distribution to a process $\{\overline{Q}(\cdot), \overline{F}(\cdot), \overline{\Pi}(\cdot)\}$, which has the following (easily verifiable) structure and properties:

(a) $\overline{Q}(\cdot)$ is stationary, with flux equal $\lambda$;
(b) $\overline{Q}_n(t) \leq 1$ for all $n$ and all $t$;
(c) The movement of customers between sites is driven by independent, unit rate Poisson processes $\overline{\Pi}_i(\cdot)$, according to BP algorithm rules;
(d) By (b) and (c), $\overline{Q}(\cdot)$ is a TASEP in the sense that there is at most one particle at each site at any time, and the rules governing the movement of particles between sites are same as those described in Sect. 4. However, it is a "one-sided" process: it "lives" on the positive integers (as opposed to all integers). New particles may be added at ("arrive to") site 1 when it is empty; this arrival process at site 1 is a stationary process, *not* independent of the "rest of the process" $\overline{Q}(\cdot)$. Finally, $\overline{Q}(\cdot)$ is *not* Markov.

Consider the following projection of process $\overline{Q}(\cdot)$. All particles arriving at site 1 after time 0 and the particle located at site 1 at time 0 (if any), we will call "new" particles, while all particles initially present at sites $n \geq 2$ are "old". Let $Q^*(\cdot)$ denote the process "keeping track" of new particles in $\overline{Q}(\cdot)$, namely $Q^*_n(t) = 1$ if there is a new particle located at site $n$ at time $t$, and $Q^*_n(t) = 0$ otherwise. The flux of process $Q^*(\cdot)$ from site 1 to site 2 is obviously equal to the flux of $\overline{Q}(\cdot)$, which is $\lambda$. We will compare $Q^*(\cdot)$ to the following TASEP $Q^{**}(\cdot)$, coupled to it—with the same Poisson processes driving movement between sites. The initial state of $Q^{**}(\cdot)$ is: $Q^{**}_1(0) = 1$ and $Q^{**}_n(0) = 0$ for $n \geq 2$. By definition, $Q^{**}_1(t) \equiv 1$, i.e., if at any time a particle moves from site 1 to 2, another particle is immediately added at site 1. Using the argument analogous to that in the proof of Lemma 4, it is easy to see that $Q^*(t) \preceq Q^{**}(t)$, which implies

$$F^*_2(t) \leq F^{**}_2(t), \quad t \geq 0,$$

where $F^*_2(t)$ and $F^{**}_2(t)$ are the numbers of particle arrivals in $(0, t]$ at site 2 in the processes $Q^*(\cdot)$ and $Q^{**}(\cdot)$, respectively. Then,

$$\liminf_{t\to\infty} \frac{\mathbb{E}F^{**}_2(t)}{t} \geq \lim_{t\to\infty} \frac{\mathbb{E}F^*_2(t)}{t} = \lambda. \tag{17}$$

The TASEP $Q^{**}(\cdot)$ is a special case of the one-sided asymmetric simple exclusion process, studied in [7]. It is known (see Theorem 1.8(a) and Theorem 1.7(b) of [7])

that the distribution of $Q^{**}(t)$ converges to a stationary distribution, with the corresponding stationary process having flux $1/4$. This means that $\lim_{t\to\infty} \mathbb{E}F_2^{**}(t)/t = 1/4$, which contradicts (17) since $\lambda > 1/4$. The proof is complete. □

## 6 Remark on more general input processes

The Poisson assumption on the input process is adopted to simplify the exposition. We believe our main results can be easily generalized for the case of a stationary ergodic input process $A(\cdot)$, as long as the large deviations bound

$$\mathbb{P}\big\{\big|A(0,t) - \lambda t\big| > \delta t\big\} \le C_3 e^{-C_4 t}, \tag{18}$$

analogous to (11), holds for any $\delta > 0$. Moreover, if (18) does not hold, and we only have the ergodicity of $A(\cdot)$, the uniform stochastic boundedness results of Theorem 10 (and then Theorem 1(i)) and Theorem 11 will still hold, except the bounds are proper (finite w.p.1) random variables, not necessarily with exponential tails.

## 7 Conclusions

We have shown that the queues in the tandem system remain stochastically bounded [resp., grow to infinity] as the number of queues $N \to \infty$, if the system load (exogenous input rate) $\lambda$ is strictly below [resp., strictly above] the critical value $1/4$. Our analysis essentially reduces the problem to establishing stability (or instability) of the infinite-tandem system, where stability is understood as the existence of a proper stationary distribution (with all queues finite w.p.1). In the subcritical case, we construct a process providing an upper bound on the queues in the infinite system; the construction involves a stationary TASEP—a classical, well-studied process in interacting particle systems. Our treatment of the supercritical case is less constructive: we show that a stable process cannot have a flux greater than $1/4$, which is the maximum possible flux of a TASEP; this proves instability for $\lambda > 1/4$.

Note that, although it is very natural to expect that in the critical case, $\lambda = 1/4$, the infinite system is unstable, proving this fact remains an open problem. Our treatment of the subcritical case does *not* give a suitable *lower* bound on the queues, which (if grows to infinity as $\lambda \uparrow 1/4$) would prove instability for $\lambda = 1/4$ (and then for $\lambda > 1/4$), by monotonicity on $\lambda$. (That is why we used a different argument, as explained above, to prove instability for $\lambda > 1/4$.)

**Acknowledgements** I would like to thank Yuliy Baryshnikov for extremely helpful discussions throughout the course of this work. Comments from the referees and AE were very useful in improving the exposition.

## Appendix A: Proof of Lemma 3

Consider the constructive representation of the process, given in Sect. 3. Fix arbitrary $t_1 > 0$ and arbitrary $n'' > n$. W.p.1 there exists finite $n' > n''$ such that there are no

points of process $\Pi_{n'}$ in $[0, t_1]$, and the realization in $[0, t_1]$ of the process, restricted to sites up to $n'$, depends only on the realizations of $\Pi_1, \ldots, \Pi_{n'-1}$ and of the exogenous arrival process. This realization is such that, in $[0, t_1]$, a transition from a state satisfying

$$Q_n + 1 \geq \max_{n<k\leq n'} Q_k$$

to a state violating this condition is impossible, because this would violate the BP rule. Since $n''$ and $t_1$ are arbitrary, (4) holds w.p.1 for all $t \geq 0$.

## Appendix B: Proof of Lemma 4

Consider the two processes coupled on the common probability space, using the construction given in Sect. 3. Recall that the relation $Q(0) \preceq Q'(0)$ implies that $\|Q(0)\|$ and $\|Q'(0)\|$ are finite, and then w.p.1 $\|Q(t)\|$ and $\|Q'(t)\|$ are finite for all $t$, and then w.p.1 there is only a finite number of transitions of the coupled process $(Q(t), Q'(t))$ in any finite time interval. It is easy to check that a transition from state such that $Q \preceq Q'$ to a state violating this condition is impossible, because this would violate the BP rule. This proves $Q(t) \preceq Q'(t)$ for all $t$.

The proof of $Q(t) \leq Q'(t)$ is analogous to that of Lemma 3. Here, as in that proof, since $\|Q(t)\|$ and $\|Q'(t)\|$ may be infinite, we need to use a "localization in space" as an intermediate step.

The monotonicity of $Q^{(N)}(\cdot)$ is proved analogously.

## Appendix C: Proof of Proposition 5

(i) Consider all processes coupled on the common probability space, using the construction given in Sect. 3. W.p.1 there is only a finite number of transitions in any finite time interval. A transition at time $t$ cannot lead to a state violating (5) if the condition held at $t-$, which proves (5). Also, w.p.1, for any fixed $t \geq 0$, $Q^{(N)}(t) = Q(t)$ for all sufficiently large $N$, which proves $Q^{(N)}(t) \uparrow Q(t)$.

(ii) For any $t_1 \leq t_2$, we have $Q(t_1) \leq_{\mathrm{st}} Q(t_2)$ because $Q(0) \leq_{\mathrm{st}} Q(t_2 - t_1)$ and process $Q(\cdot)$ is monotone (Lemma 4). Similarly, $Q^{(N)}(t_1) \leq_{\mathrm{st}} Q^{(N)}(t_2)$. The fact that $Q^{(N_1)}(t) \leq_{\mathrm{st}} Q^{(N_2)}(t)$ for any $t$ and $N_1 \leq N_2$ follows from (i).

(iii) Convergence (6) follows from the ergodicity of the countable Markov chain $Q^{(N)}(\cdot)$. Property $Q^{(N)}(\infty) \leq_{\mathrm{st}} Q^{(N+1)}(\infty)$ is equivalent to $\mathbb{E}f[Q^{(N)}(\infty)] \leq \mathbb{E}f[Q^{(N+1)}(\infty)]$ for any monotone (w.r.t. $\leq$ order) continuous function $f$; the latter property in turn follows from $\mathbb{E}f[Q^{(N)}(t)] \leq \mathbb{E}f[Q^{(N+1)}(t)]$ and (6). Convergence (7) follows from the fact that a stochastically increasing sequence of random elements in a compact space must converge in distribution to a unique (in distribution) random element. To prove (8), denote by $Q'(\infty)$ the unique limit (in distribution) of $Q(t)$ as $t \to \infty$. (Limits in distribution exist because $\bar{Z}_+^S$ is compact; the uniqueness follows form $Q(t)$ being stochastically increasing.) For a monotone continuous function $f$, we have, for any $N$:

$$\mathbb{E}f\big[Q'(\infty)\big] = \lim_t \mathbb{E}f\big[Q(t)\big] \geq \lim_t \mathbb{E}f\big[Q^{(N)}(t)\big] = \mathbb{E}f\big[Q^{(N)}(\infty)\big]$$

and therefore $Q(\infty) \leq_{\mathrm{st}} Q'(\infty)$; on the other hand,

$$\mathbb{E}f\big[Q(t)\big] = \lim_N \mathbb{E}f\big[Q^{(N)}(t)\big] \leq \lim_N \mathbb{E}f\big[Q^{(N)}(\infty)\big] = \mathbb{E}f\big[Q(\infty)\big]$$

and therefore $Q'(\infty) \leq_{\mathrm{st}} Q(\infty)$. Thus, $Q'(\infty) =_{\mathrm{st}} Q(\infty)$.

(iv) The stationarity of the distribution of $Q(\infty)$ follows from (8) and Proposition I.1.8(d) in [8]. By the monotonicity of process $Q(\cdot)$, any stationary distribution of $Q(\cdot)$ must stochastically dominate the limit of the distribution of $Q(t)$ with $Q(0)$ being the zero (empty) state.

(v) Obviously, condition (4 holds for $t = 0$, and then for all $t \geq 0$ by Lemma 3. From convergence (8), for any fixed $n$ and $n' > n$, w.p.1 we have

$$Q_n(t) + 1 \geq \max_{n<k\leq n'} Q_k(t),$$

which implies (4).

(vi) Follows from (ii) and (8).

## Appendix D: Proof of Proposition 6

(i) By the construction of the process, $Q_n(t)$ remains finite w.p.1 for all finite $t$. Then, by the BP rule, the input flow into queue $n$ is the unit rate Poisson process. If we assume that a customer is removed (served) from queue $n$ at every point of $\Pi_n$, we obtain process $Q'_n(t)$ such that $Q'_n(t) \leq Q_n(t)$. But, the evolution of $Q'_n(t)$ is that of the queue in a critically loaded M/M/1 system, which implies that $Q'_n(t) \to \infty$, and then $Q_n(t) \to \infty$, in probability as $t \to \infty$.

(ii) Suppose $\mathbb{P}\{Q_1(\infty) = \infty\} = \delta > 0$. Let us show that $\delta = 1$. By the monotonicity of $Q(\cdot)$, uniformly on all fixed initial states $Q(0)$, for any finite $C > 0$,

$$\liminf_{t\to\infty} \mathbb{P}\big\{Q_1(t) \geq C\big\} \geq \delta.$$

By the process construction, if $Q_1(0) = \infty$ then $Q_1(t) = \infty$ for all $t$. Consider a stationary version of $Q(\cdot)$ with stationary state distributed as $Q(\infty)$. Then, given $Q_1(0)$ is infinite or finite with probabilities $\delta$ and $1 - \delta$, respectively,

$$\mathbb{P}\big\{Q_1(\infty) \geq C\big\} = \liminf_{t\to\infty} \mathbb{P}\big\{Q_1(t) \geq C\big\} \geq \delta \cdot 1 + (1 - \delta)\delta.$$

Since $C$ is arbitrary, this means $\delta = \mathbb{P}\{Q_1(\infty) = \infty\} \geq \delta + (1 - \delta)\delta$, which implies that $\delta = 1$. Thus, $Q_1(\infty) = \infty$ w.p.1. Then, by statement (i), $Q_2(\infty) = \infty$ w.p.1. Repeated use of statement (i), proves $Q_n(\infty) = \infty$ w.p.1 for all $n$.

Suppose now $\mathbb{P}\{B(Q(\infty)) = \infty\} = \delta > 0$. Using the same argument as above for $Q_1(\infty)$, we show that $\delta = 1$, namely w.p.1 all $Q_n(\infty) \geq 1$. By the definition of BP rule, if the initial state is such that all queues are at least 1, and we reduce all queues by 1, the process does not change, except all queues remain smaller by 1. This means that $(Q_n(\infty) - 1, n = 1, 2, \ldots)$ is also a stationary state. But, since $Q(\infty)$ is the stochastic lower bound of any stationary state, this is only possible when $Q_n(\infty) = \infty$ w.p.1 for all $n$.

## Appendix E: Proof of Proposition 7

Consider a (potential) transition, triggered by a specific point $\tau$ of a realization $\Pi_n$. If $Q'(\tau-) \preceq Q(\tau-)$, then $Q'(\tau) \preceq Q(\tau)$. Indeed, the transition at $\tau$ can be considered in two "steps": first, the transition is as if $\tau$ is a valid point; second, if a customer in $Q'$ moved from $n$ to $n+1$, but $\tau$ happens to be invalid, we move this customer "back". In the first step, $Q'(\tau) \preceq Q(\tau)$ is preserved by Lemma 4, and in the second step—by definition of $\preceq$ order. Finally, $Q'(\tau) \preceq Q(\tau)$ implies $Q'_1(t) \geq Q_1(t)$, because $\|Q'(t)\| = \|Q(t)\|$.

## Appendix F: Proof of Proposition 8

(i) Follows from Proposition 7.

(ii) Condition $Q'_n(t) \leq Y_n(t)$, $n \geq 2$, holds at $t = 0$, and cannot be violated by a transition at time $\tau$ if it held at $\tau-$.

(iii) By (ii), if there is a jump from 1 to 2 in TASEP $Y$ at $\tau$, then $Y_2(\tau-) = 0$, and then $Q'_2(\tau-) = 0$. The statement easily follows.

## References

1. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1968)
2. Bui, L., Srikant, R., Stolyar, A.L.: Optimal resource allocation for multicast flows in multihop wireless networks. Philos. Trans. R. Soc. Lond. A **366**, 2059–2074 (2008)
3. Bui, L., Srikant, R., Stolyar, A.L.: Novel architectures and algorithms for delay reduction in back-pressure scheduling and routing. In: Proceeding of INFOCOM'2009 Mini-conference (2009)
4. Dai, J.G., Lin, W.: Maximum pressure policies in stochastic processing networks. Oper. Res. **53**, 197–218 (2005)
5. Duffield, N.G., O'Connell, N.: Large deviations and overflow probabilities for the general single-server queue, with applications. Proc. Camb. Philos. Soc. **118**, 363–374 (1995)
6. Glynn, P.W., Whitt, W.: Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. J. Appl. Probab. A **31**, 413–430 (1994)
7. Liggett, T.M.: Ergodic theorems for the asymmetric simple exclusion process. Trans. Am. Math. Soc. **213**, 237–261 (1975)
8. Liggett, T.M.: Interacting Particle Systems. Springer, New York (1985)
9. Liggett, T.M.: Stochastic Interacting Systems: Contact, Voter and Exclusion Processes. Springer, New York (1999)
10. Loynes, R.M.: The stability of a queue with non-independent inter-arrival and service times. Proc. Camb. Philos. Soc. **58**(3), 497–520 (1962)
11. Martin, J.B.: Large tandem queueing networks with blocking. Queueing Syst. **41**, 45–72 (2002)
12. Stolyar, A.L.: Maximizing queueing network utility subject to stability: greedy primal-dual algorithm. Queueing Syst. **50**, 401–457 (2005)
13. Tassiulas, L., Ephremides, A.: Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. IEEE Trans. Autom. Control **37**, 1936–1948 (1992)