Systems & Control Letters 23 (1994) 305–314
North-Holland



# On infinite queueing tandems*

Nicholas Bambos and Balaji Prabhakar

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

Received 10 December 1992
Revised 9 August 1993

*Abstract:* We study the flow of jobs on an infinite series of first-come-first-served queues. Jobs are placed in the buffer of the first queue and allowed to flow through the infinite tandem of queues. The service times of each job on consecutive queues form a stationary and ergodic sequence. We are interested in characterizing the flow of jobs asymptotically, after they have passed through a large number of queues.

It is shown that the job flow reaches asymptotically a stationary state, which can be characterized in terms of the average service times of the jobs. They eventually form *clusters*, such that every two consecutive jobs belonging to the same cluster collide infinitely often, while jobs belonging to different clusters eventually cease to interact.

*Keywords:* Infinite series of queues; stationary tandem queueing networks; stability; stochastic flows.

## 1. Introduction

Consider an infinite series of first-come-first-served (FCFS) queues, indexed by the positive integers $k \in \mathscr{Z}_+ = \{1, 2, 3, \ldots\}$, and a set of jobs to be served at the queues. A job leaving the $k$th queue joins the $(k+1)$ st one. There are $J \in \mathscr{Z}_+$ jobs, which are initially queued up in the buffer, the first ($k = 1$) queue, while all other queues are empty. The jobs are indexed by $j \in \mathscr{J} = \{1, 2, 3, \ldots, J\}$ and are numbered according to their ordering in the buffer, i.e. job 1 starts service first, job 2 second, and so on. At time 0 we start serving the jobs, letting them flow from the first buffer through the queues, receiving service at each one of them. The buffers of all the queues have capacity at least $J$, so that no buffer overflows or blocking of jobs occur (although such cases can be studied similarly; see Remark 6). All the servers work with service rate 1. Let $\sigma_j^k$ be the service time requirement of the $j$th job on the $k$th queue. The $\sigma_j^k$'s are identified with the elements of an infinite sequence

$$\boldsymbol{S} = \{\mathscr{S}^k = \{\sigma_j^k, j \in \mathscr{J}\}, k \in \mathscr{Z}\}, \tag{1}$$

which is defined on some probability space $(\Omega, \mathscr{F}, P)$ and is assumed to be stationary and ergodic with respect to the transformation

$$\theta_v \boldsymbol{S} = \{\mathscr{S}^{k-v}, k \in \mathscr{Z}\}, \tag{2}$$

corresponding to $k$-index shifts. Additionally, we assume that $E[\sigma_j^0] < \infty$ for every $j \in \mathscr{J}$. Observe that the sequence $\boldsymbol{S}$ contains terms with $k \leqslant 0$ too, but only those with $k > 0$ are given a physical meaning in the previous setup. Also let $t_j^k$ be the departure time of the $j$th job from the $k$th queue and define the interdeparture times of the jobs exiting the $k$th queue by

$$\mathscr{T}^k = \{\tau_j^k = (t_{j+1}^k - t_j^k), j \in \mathscr{J}'\}, \tag{3}$$

where $\mathscr{J}' = \{1, 2, 3, \ldots, J-1\}$.

*Correspondence to:* N. Bambos, Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA 90024, USA.

* Research supported in part by grants NSF-DDM-RIA-9010778, NSF-NCR-9116268, NSF-NCR-NYI-9258507, by an AT & T Foundation Grant and a GTE Fellowship.


*SSDI* 0167-6911(93)E0135-4

The jobs moving through the queues constitute a flow which can be mathematically characterized by the job interdeparture times. We are interested in the asymptotic statistics of this flow, specifically, whether $\{\mathcal{T}^k, k \in \mathcal{Z}_+\}$ converges in some probabilistic sense as $k \to \infty$.

From a practical point of view, the queueing tandem described here could correspond to a long assembly line in flexible manufacturing, a long processing pipeline in a computer architecture, or a packet communication link with a large number of hops. Our original motivation for studying this system was to estimate the intercompletion times of jobs on long pipelines.

Addressing the problem of invariant measures of the G/G/1 queueing operator, Bambos and Walrand [2] made an observation (discussed in the following section), which is applicable to our situation here, proving the existence of the asymptotic flow. However, the basic issue of finiteness (proper distributions) of this flow was not resolved, and no such flow characterization was established. Concerning finite queueing tandems, we have the seminal work of Loynes [6] on the stability theory of G/G/1 queues and tandem networks of them with stationary interarrival and service times (also see [7]). Earlier work of Lindley [5] has provided structural and computational results on queues with i.i.d. interarrival and service times. Glynn and Whitt [4] have lately studied large queueing tandems with i.i.d. service times, obtaining results of the central-limit-theorem type, in the critical case where all jobs have the same service statistics. The static tandem network problem, concerned with the computation of the asymptotic average execution time per job on a finite queueing tandem, as the number of jobs in the front buffer tends to infinity, has been addressed by Bambos and Walrand [1] under stationary service times.

The objective of this work is to characterize the asymptotic flow in terms of finiteness (proper distributions of interdeparture times), contributing to the theory of infinite queueing tandems, which is related to smoothing properties and invariant measures of queueing operators. This is done under stationary ergodic service times, which are essentially the most general. Theorem 1 provides the necessary characterization in terms of the expected service times of jobs. As explained in Remark 2, the jobs eventually form stationary clusters, each traveling at constant speed. Each cluster consists of consecutive jobs that collide (catch up) infinitely often, while their distances (interdeparture times) asymptotically reach proper distributions. For some special cases a dual network structure provides insight into the results (Remark 3). The application of the results to invariant measures of queueing operators is also discussed (Remark 4).

## 2. The asymptotic flow

We consider each queue as a random operator that maps its interarrival times to interdeparture ones, taking the point of view of Bambos and Walrand [2]. Letting $w_j^k$ be the waiting time of the $j$th job at the $k$th queue (between arriving and starting being served there), we get

$$w_{j+1}^k = [w_j^k + \sigma_j^k - \tau_j^{k-1}]^+, \tag{4}$$

($[x]^+ = xI_{\{x \geq 0\}}$), where $w_1^k = 0$ for every $k \in \mathcal{Z}$. Moreover, since the interdeparture times at the $(k-1)$th queue are the interarrival times of the $k$th one, we get $\tau_j^k = t_{j+1}^k - t_j^k = t_{j+1}^{k-1} + w_{j+1}^k + \sigma_{j+1}^k - (t_j^{k-1} + w_j^k + \sigma_j^k) = \tau_j^{k-1} + [w_j^k + \sigma_j^k - \tau_j^{k-1}]^+ + \sigma_{j+1}^k - w_j^k - \sigma_j^k$, which leads to (by manipulating the $[\ ]^+$)

$$\tau_j^k = [\tau_j^{k-1} - \sigma_j^k - w_j^k]^+ + \sigma_{j+1}^k \tag{5}$$

for every $j \in \mathcal{J}'$, $k \in \mathcal{Z}_+$. Therefore, there is a well-defined *deterministic* function $\boldsymbol{F}: \mathcal{R}^{\mathcal{J}} \times \mathcal{R}^{\mathcal{J}'} \to \mathcal{R}^{\mathcal{J}'}$ mapping sequences of interarrival and service times of a queue to sequences of interdeparture times, i.e.

$$\mathcal{T}^k = \boldsymbol{F}(\mathcal{S}^k, \mathcal{T}^{k-1}). \tag{6}$$

Consequently, for the $k$th queue, we can define the random mapping $\boldsymbol{F}^k: \mathcal{R}^{\mathcal{J}'} \to \mathcal{R}^{\mathcal{J}'}$ by

$$\mathcal{T}^k = \boldsymbol{F}^k(\mathcal{T}^{k-1}) = \boldsymbol{F}(\mathcal{S}^k, \mathcal{T}^{k-1}), \tag{7}$$

absorbing the stochasticity of the service times $\mathscr{S}^k$ into the functional form. We can then write

$$\mathscr{T}^k = [\boldsymbol{F}^k \circ \boldsymbol{F}^{k-1} \circ \boldsymbol{F}^{k-2} \circ \cdots \circ \boldsymbol{F}^2 \circ \boldsymbol{F}^1](\mathscr{T}^0 = \vec{0}), \tag{8}$$

where the symbol $\circ$ denotes composition of mappings. We are interested in characterizing the asymptotic distribution of $\mathscr{T}^k$ as $k \to \infty$.

Because of the tandem structure of the queueing network, the departure time of the $j$th job from the $k$th queue can be expressed directly as a function of the service times by

$$t_j^k = \max_{1 \leqslant m_1 \leqslant m_2 \leqslant m_3 \leqslant \cdots \leqslant m_{k-1} \leqslant j} \left\{ \sum_{\mu=1}^{m_1} \sigma_\mu^1 + \sum_{\mu=m_1}^{m_2} \sigma_\mu^2 + \sum_{\mu=m_2}^{m_3} \sigma_\mu^3 + \cdots + \sum_{\mu=m_{k-1}}^{j} \sigma_\mu^k \right\} \tag{9}$$

or, equivalently, by

$$t_j^k = \max_{1 \leqslant n_1 \leqslant n_2 \leqslant n_3 \leqslant \cdots \leqslant n_{j-1} \leqslant k} \left\{ \sum_{\nu=1}^{n_1} \sigma_1^\nu + \sum_{\nu=n_1}^{n_2} \sigma_2^\nu + \sum_{\nu=n_2}^{n_3} \sigma_3^\nu + \cdots + \sum_{\nu=n_{j-1}}^{k} \sigma_j^\nu \right\}, \tag{10}$$

for every $j \in \mathscr{J}$, $k \in \mathscr{Z}_+$ (this can be proven inductively; see [3] for a proof). Therefore, the interdeparture times $\tau_j^k = t_{j+1}^k - t_j^k$ can be analytically expressed using the previous relations. We will use this expression in the proof of the theorem later.

The original tandem of queues can be (conceptually) extended, by appending an additional infinite tandem of queues (indexed by $k \leqslant 0$) on the left of the original one, creating a sequence of queues in series, ranging from $-\infty$ to $+\infty$. Again, jobs leaving the $k$th queue join the $(k+1)$st one. The service time of the $j$th job ($j \in \mathscr{J}$) on the $k$th queue ($k \in \mathscr{Z}$) is the $\sigma_j^k$ element of the sequence $\boldsymbol{S}$, which already provides for the above-mentioned extension by including elements with negative $k$-indices. The set of random mappings $\boldsymbol{F}^k$'s can also be naturally extended over all integers $\{\boldsymbol{F}^k, k \in \mathscr{Z}\}$.

On the extended queueing structure (doubly infinite tandem) we define pathwise the quantity

$$\boldsymbol{\Phi}^{n,m} = \{\phi_j^{n,m}, j \in \mathscr{J}'\} = [\boldsymbol{F}^n \circ \boldsymbol{F}^{n-1} \circ \boldsymbol{F}^{n-2} \circ \cdots \circ \boldsymbol{F}^{m+1} \circ \boldsymbol{F}^m](\vec{0}), \tag{11}$$

where $m < n$ and $m, n \in \mathscr{Z}$. Note that $\phi_j^{n,m}$ is the interdeparture time between the $(j+1)$st job and the $j$th one at the $n$th queue, given that the interarrival times of all the jobs at the $m$th queue were all zero (all jobs entered concurrently in the buffer of the $m$th queue). It is easy to see that for every $k \in \mathscr{Z}_+$ we have

$$\mathscr{T}^k = \boldsymbol{\Phi}^{k,1}. \tag{12}$$

Working on a related problem, Bambos and Walrand [2] observed that for every $j \in \mathscr{J}'$ the quantity $\phi_j^{n,m}$ is increasing as $m \to -\infty$ ($j, n$ fixed); thus the limits

$$\phi_j^n = \lim_{m \to -\infty} \phi_j^{n,m} \tag{13}$$

exist pathwise for every $n \in \mathscr{Z}$, and similarly we define $\boldsymbol{\Phi}^n = \{\phi_j^n, j \in \mathscr{J}'\} = \lim_{m \to -\infty} \boldsymbol{\Phi}^{n,m}$. However, it has not been possible to prove whether these limits are finite or infinite almost surely, which is a basic issue in our study of the asymptotic flow. This matter is settled in the following theorem, which also characterizes the asymptotic flow in terms of the expected processing times of the jobs.

In preparation for the proof let us define the notion of *collision* between two jobs. We say that during the evolution of the system the $(j+1)$st job collides with the $j$th one on the $k$th queue if the former has to wait for the latter to exit the queue before it can start service there, i.e. the $(j+1)$st job *catches up* with the $j$th one. It is easy to see that, on a given sample path, we have

$$\tau_j^k = \sigma_{j+1}^k \tag{14}$$

if the two jobs collide at the $k$th queue (of course $\tau_j^k \geqslant \sigma_{j+1}^k$ always).

**Theorem 1.** *For any $\theta$-stationary and ergodic sequence of service times $\boldsymbol{S}$ we have the following:*

(a) *For any $j \in \mathscr{J}$, if $E[\sigma_{j+1}^0] < \max_{\{1 \leqslant i \leqslant j\}} \{E[\sigma_i^0]\}$, then*

$$\phi_j^n < \infty \tag{15}$$

*almost surely for every* $n \in \mathscr{L}$,

$$\lim_{k\to\infty} P[\tau_j^k \leqslant x] = P[\phi_j^0 \leqslant x] \tag{16}$$

*for every* $x \in [0, \infty)$, *and the* $(j+1)$*st job* collides (*catches up*) *with the* $j$*th one* infinitely often *almost surely.*

(b) *For any* $j \in \mathscr{J}$, *if* $E[\sigma_{j+1}^0] > \max_{\{1 \leqslant i \leqslant j\}} \{E[\sigma_i^0]\}$, *then*

$$\phi_j^n = \infty \tag{17}$$

*almost surely for every* $n \in \mathscr{L}$,

$$\lim_{k\to\infty} \left[\frac{\tau_j^k}{k}\right] = E[\sigma_{j+1}^0] - \max_{1 \leqslant i \leqslant j} \{E[\sigma_i^0]\} > 0, \tag{18}$$

*and the* $(j+1)$*st job* collides *with the* $j$*th one only* finitely often *almost surely.*

**Proof.** (a) We start with some definitions and observations. Given that all jobs are concurrently placed in the buffer of the $m$th queue (interarrival times there are zero), define $c_a^{n,m}$ to be the *last queue*, before the $n$th one ($m < n, c_a^{n,m} < n$), where the $(a+1)$th job collides with the $a$th one ($a \in \mathscr{J}'$). This implies that

$$\phi_a^{c_a^{n,m},m} = \sigma_{a+1}^{c_a^{n,m}} \tag{19}$$

by (14), and the $(a+1)$st job never waits on the queues between the $(c_a^{n,m})$th and $n$th ones. Defining $D_a^{n,m}$ to be the departure time of the $a$th job from the $n$th queue (given that the jobs entered concurrently at the $m$th one), we get

$$D_{a+1}^{n,m} = D_a^{n,m} + \phi_a^{n,m}. \tag{20}$$

Since the $(a+1)$st job never waits at queues $\{c_a^{n,m}+1, c_a^{n,m}+2, \ldots, -1, n\}$, we have

$$D_{a+1}^{n,m} = D_{a+1}^{c_a^{n,m},m} + \sum_{v=c_a^{n,m}+1}^{n} \sigma_{a+1}^v = D_a^{c_a^{n,m},m} + \sigma_{a+1}^{c_a^{n,m}} + \sum_{v=c_a^{n,m}+1}^{n} \sigma_{a+1}^v \tag{21}$$

for every $m < n$. Moreover, for every $m < n$, $a \in \mathscr{J}'$ and $b \in \{1, 2, 3, \ldots, a-1, a\}$, we have

$$D_a^{n,m} - D_a^{c_a^{n,m},m} \geqslant \sum_{v=c_a^{n,m}+1}^{n} \sigma_b^v - (D_a^{c_a^{n,m},m} - D_b^{c_a^{n,m},m}), \tag{22}$$

since the $b$th job has to exit the $n$th queue before the $a$th job does, and the former can exit the $n$th queue in time at least $\sum_{v=c_a^{n,m}+1}^{n} \sigma_b^v + D_b^{c_a^{n,m},m}$ (minimum achieved when no collision occurs after the $(c_a^{m,n})$th queue). Now, by $D_a^{c_a^{n,m},m} - D_b^{c_a^{n,m},m} = \sum_{\rho=b}^{a-1} (D_{\rho+1}^{c_a^{n,m},m} - D_\rho^{c_a^{n,m},m}) = \sum_{\rho=b}^{a-1} \phi_\rho^{c_a^{n,m},m}$, we get

$$D_a^{n,m} - D_a^{c_a^{n,m},m} \geqslant \sum_{v=c_a^{n,m}+1}^{n} \sigma_b^v - \sum_{\rho=b}^{a-1} \phi_\rho^{c_a^{n,m},m}. \tag{23}$$

Finally, using (20) and (21), we get

$$\phi_a^{n,m} = D_{a+1}^{n,m} - D_a^{n,m} = D_a^{c_a^{n,m},m} + \sum_{v=c_a^{n,m}}^{n} \sigma_{a+1}^v - D_a^{n,m} = \sum_{v=c_a^{n,m}}^{n} \sigma_{a+1}^v - (D_a^{n,m} - D_a^{c_a^{n,m},m}), \tag{24}$$

and by (23) we have (pathwise, for every $m < n$, $a \in \mathscr{J}'$, $b \in \{1, 2, \ldots, a-1, a\}$)

$$\phi_a^{n,m} \leqslant \sum_{v=c_a^{n,m}}^{n} \sigma_{a+1}^v - \sum_{v=c_a^{n,m}+1}^{n} \sigma_b^v + \sum_{\rho=b}^{a-1} \phi_\rho^{c_a^{n,m},m}, \tag{25}$$

which is a *basic structural relation* used in the proof below.

Now, since $\phi_a^{n,m}$ increases as $m \to -\infty$, the $c_a^{n,m}$ is decreasing as $m \to -\infty$ pathwise ($a, n$ fixed). Therefore, define

$$C_a^n = \lim_{m \to -\infty} [c_a^{n,m}] \tag{26}$$

(exists almost surely, but may be finite or $-\infty$) for every $a \in \mathscr{J}'$. Finally, define $j^* = \max\{a \leqslant j : E[\sigma_a^0] > E[\sigma_{j+1}^0]\}$ to be the index of the job closest to (but preceding) the $(j+1)$st one with average service time larger than that of the latter. Therefore,

$$E[\sigma_b^0] < E[\sigma_{j^*}^0] \quad \text{for every } b \in \{j^*+1, \ldots, j, j+1\}. \tag{27}$$

By our assumption that $E[\sigma_{j+1}^0] < \max_{\{1 \leqslant i \leqslant j\}} \{E[\sigma_i^0]\}$, it is guaranteed that such a $j^* \leqslant j$ exists.

We now prove that $\phi_j^n < \infty$, using *induction* on the index set $\{j^*, j^*+1, \ldots, j-1, j\}$. This is done in two steps, each having three key substeps.

*Step* 1.1: To prove $\phi_{j^*}^n < \infty$, we need to show that $\lim_{m \to \infty} [c_{j^*}^{n,m}] = C_{j^*}^n > -\infty$ (is finite). Arguing by contradiction suppose that $C_{j^*}^n = -\infty$. From the structural relation (25), taking $a = b = j^*$, we get $0 \leqslant \phi_{j^*}^{n,m} \leqslant \sum_{v=c_{j^*}^{n,m}}^{n} \sigma_{j^*+1}^v - \sum_{v=c_{j^*}^{n,m}+1}^{n,m} \sigma_{j^*}^{n,m}$, pathwise. Dividing by $n - c_{j^*}^{n,m}$, letting $m \to -\infty$, and using Birkhoff's ergodic theorem [8] on the stationary sequences $\{\sigma_{j^*+1}^v, v \in \mathscr{Z}\}$ and $\{\sigma_{j^*}^v, v \in \mathscr{Z}\}$, we get

$$0 \leqslant \lim_{m \to -\infty} \frac{\sum_{v=c_{j^*}^{n,m}}^{n} \sigma_{j^*+1}^v}{n - c_{j^*}^{n,m}} - \lim_{m \to -\infty} \frac{\sum_{v=c_{j^*}^{n,m}+1}^{n} \sigma_{j^*}^v}{n - c_{j^*}^{n,m}} = E[\sigma_{j^*+1}^0] - E[\sigma_{j^*}^0], \tag{28}$$

which is a contradiction, because of (27). Therefore, $\lim_{m \to -\infty} [c_{j^*}^{n,m}] = C_{j^*}^n > -\infty$ (is finite), and taking the limits for $m \to -\infty$, we get $0 \leqslant \phi_{j^*}^n = \lim_{m \to -\infty} [\phi_{j^*}^{n,m}] \leqslant \sum_{v=C_{j^*}^n}^{n} \sigma_{j^*+1}^v - \sum_{v=C_{j^*}^n+1}^{n} \sigma_{j^*}^v < \infty$, almost surely, for every $n \in \mathscr{Z}$, as required.

Following the same reasoning as above, we can prove that at the limit $m \to -\infty$ (where $\{\phi_{j^*}^v\}$ is attained) the $(j^*+1)$st job collides with the $(j^*)$th one *infinitely often* before queue $n$. Indeed, $C_{j^*}^n$ is the 'collision point' (finite) immediately before queue $n$ (arbitrary) in $\{\phi_{j^*}^v, v \in \mathscr{Z}\}$. Choosing queue $C_{j^*}^n$, instead of $n$, to apply the arguments of the previous paragraph, we construct the collision point preceding $C_{j^*}^n$, and so on. Therefore, there is a strictly increasing infinite sequence $\{C_{j^*}(k)\}$ with elements that are the indices of the queues where the above-mentioned jobs collide in $\{\phi_{j^*}^v, v \in \mathscr{Z}\}$. The $k$-numbering is so chosen that $-\infty < \cdots C_{j^*}(-1) < C_{j^*}(0) \leqslant 0 < C_{j^*}(1) < C_{j^*}(2) \cdots < n$ and we have $\lim_{k \to -\infty} [C_{j^*}(k)] = -\infty$. Since the two jobs collide only at the queues $C_{j^*}(k)$, we have $\phi_{j^*}^{C_{j^*}(k)} = \sigma_{j^*+1}^{C_{j^*}(k)}$, while $\phi_{j^*}^v > \sigma_{j^*+1}^v$ when $C_{j^*}(k) < v < C_{j^*}(k+1)$.

*Step* 1.2: The second point of the basic induction step is that

$$\lim_{k \to -\infty} \frac{C_{j^*}(k+1)}{C_{j^*}(k)} = 1 \tag{29}$$

almost surely, which we use later. Indeed, arguing by contradiction suppose that, on a sample path, we have $\liminf_{k \to -\infty} C_{j^*}(k+1)/C_{j^*}(k) = 1 - \delta$ for some positive $\delta > 0$ (note that $C_{j^*}(k) < C_{j^*}(k+1) < 0$ as $k \to -\infty$). Arguing for $\{\phi_{j^*}^n\}$ as in (25), with $a = b = j^*$, we get $0 \leqslant \phi_{j^*}^{C_{j^*}(k+1)-1} \leqslant \sum_{v=C_{j^*}(k)}^{C_{j^*}(k+1)-1} \sigma_{j^*+1}^v - \sum_{v=C_{j^*}(k)+1}^{C_{j^*}(k+1)-1} \sigma_{j^*}^v$ for every $k \in \mathscr{Z}$, and dividing by $-C_{j^*}(k) > 0$ and rearranging the terms we get

$$0 \leqslant \left( \frac{\sum_{v=C_{j^*}(k)}^{0} \sigma_{j^*+1}^v}{-C_{j^*}(k)} - \frac{\sum_{v=C_{j^*}(k+1)}^{0} \sigma_{j^*+1}^v}{-C_{j^*}(k+1)} \times \frac{C_{j^*}(k+1)}{C_{j^*}(k)} \right) \tag{30}$$

$$- \left( \frac{\sum_{v=C_{j^*}(k)+1}^{0} \sigma_{j^*}^v}{-C_{j^*}(k)} - \frac{\sum_{v=C_{j^*}(k+1)}^{0} \sigma_{j^*}^v}{-C_{j^*}(k+1)} \times \frac{C_{j^*}(k+1)}{C_{j^*}(k)} \right) \tag{31}$$

for every $k < 0$. Letting $k \to -\infty$ on a subsequence on which the $\liminf 1 - \delta$ is attained, and using Birkhoff's ergodic theorem [8], we get

$$0 \leqslant (E[\sigma^0_{j^*+1}] - E[\sigma^0_{j^*+1}](1-\delta)) - (E[\sigma^0_{j^*}] - E[\sigma^0_{j^*}](1-\delta)) = (E[\sigma^0_{j^*+1}] - (E[\sigma^0_{j^*}]) \times \delta < 0, \quad (32)$$

which leads to a contradiction for every positive $\delta$, because of (27), proving (29).

*Step* 1.3: The final point is that, using (29), we can now show that

$$\lim_{n \to -\infty} \frac{\phi^n_{j^*}}{n} = 0 \quad (33)$$

almost surely. Indeed, let $k_n = \{k \in \mathscr{Z}: C_{j^*}(k) < n \leqslant C_{j^*}(k+1)\}$ and observe that $0 \leqslant \phi^n_{j^*} \leqslant \sum_{v=C_{j^*}(k_n)}^{C_{j^*}(k_n+1)} \sigma^v_{j^*+1}$ pathwise. For negative $n$'s, dividing by $-n > 0$, noting that $-C_{j^*}(k_n+1) \leqslant -n$ and rearranging the terms, we get

$$0 \leqslant \frac{\phi^n_{j^*}}{-n} \leqslant \frac{\sum_{v=C_{j^*}(k_n)}^{C_{j^*}(k_n+1)} \sigma^v_{j^*+1}}{-C_{j^*}(k_n+1)} \leqslant \frac{\sum_{v=C_{j^*}(k_n)}^{0} \sigma^v_{j^*+1}}{-C_{j^*}(k_n)} \times \frac{C_{j^*}(k_n)}{C_{j^*}(k_n+1)} - \frac{\sum_{v=C_{j^*}(k_n+1)+1}^{0} \sigma^v_{j^*+1}}{-C_{j^*}(k_n+1)}. \quad (34)$$

Taking the limits as $n \to -\infty$, and using (29) and Birkhoff's ergodic theorem, we get $0 \leqslant \lim_{n \to -\infty} \phi^n_{j^*}/-n \leqslant E[\sigma^0_{j^*}] \times 1 - E[\sigma^0_{j^*}] = 0$ almost surely, and (33) is proven.

*Step* 2. Having proven (29) and (33), which form the first step of our inductive proof, we now assume that the following are true:

$$\phi^n_\rho < \infty \quad (35)$$

and

$$\lim_{n \to -\infty} \frac{\phi^n_\rho}{n} = 0 \quad (36)$$

for every $\rho \in \{j^*, j^*+1, \ldots, a-1\}$, where $a < j$. It is then enough to prove that (35) and (36) are true for $\rho = a$, in order to establish the validity of $\phi^n_j < \infty$ by induction.

Using the basic structural relation (25) with $b = j^*$, we get

$$0 \leqslant \phi^{n,m}_a \leqslant \sum_{v=c^{n,m}_a}^{n} \sigma^v_{a+1} - \sum_{v=c^{n,m}_a+1}^{n} \phi^v_{j^*} + \sum_{\rho=j^*}^{a-1} \phi^{c^{n,m}_a, m}_\rho. \quad (37)$$

Arguing as in step 1.1, we see that if $C^n_a = \lim_{m \to -\infty} [c^{n,m}_a] = -\infty$, then, dividing (37) by $n - c^{n,m}_a$, letting $m \to -\infty$, and using (35) and (36), we get $0 \leqslant E[\sigma^0_{a+1}] - E[\sigma^0_{j^*}] < 0$, which is a contradiction because of (27). Therefore, $C^n_a > -\infty$ (step 2.1) and, letting $m \to -\infty$ in (37), we get $0 \leqslant \phi^n_a \leqslant \sum_{v=C^n_a}^{n} \sigma^v_{a+1} - \sum_{v=C^n_a+1}^{n} \phi^v_{j^*} + \sum_{\rho=j^*}^{a-1} \sigma^{C^n_a}_\rho < \infty$, as required.

We still have to prove that $\lim_{n \to -\infty} \phi^n_a/n = 0$ (step 2.3). The proof is analogous to that of (33). First, we prove that $\lim_{k \to -\infty} C_a(k+1)/C_a(k) = 1$ (step 2.2), where $C_a(k)$ are the consecutive queues on which the $(a+1)$st job collides with the $a$th job. Reasoning on $\{\phi^n_a\}$ as in (37), we get $0 \leqslant \phi^{C_a(k+1)-1}_a \leqslant \sum_{v=C_a(k)}^{C_a(k+1)-1} \sigma^v_{a+1} - \sum_{v=C_a(k)+1}^{C_a(k+1)-1} \sigma^v_{j^*} + \sum_{\rho=j^*}^{a-1} \phi^{C_a(k)}_\rho$; the new aspect here, compared to the proof of (29), is that the rightmost part of the previous expression includes the additional term $\sum_{\rho=j^*}^{a-1} \phi^{C_a(k)}_\rho$. However, due to (36), when we divide by $C_a(k)$ and let $k \to -\infty$, this term is squeezed to zero (that is why we had to go through (33), (35) and (36) before) and the situation becomes completely analogous to (30), (31). This completes the induction argument, which can then be directly applied for $a = j$ to prove $\phi^n_j < \infty$ for all $n \in \mathscr{Z}$.

To prove (16), note that pathwise $\tau^k_j(\boldsymbol{S}) = \phi^{k,1}_j(\boldsymbol{S}) = \phi^{0,1-k}_j(\theta_k \boldsymbol{S})$ for every $k \in \mathscr{Z}_+$. Due to the $\theta$-stationarity of $\boldsymbol{S}$, we get $P[\tau^k_j(\boldsymbol{S}) \leqslant x] = P[\phi^{0,1-k}_j(\theta_k \boldsymbol{S}) \leqslant x] = P[\phi^{0,1-k}_j(\boldsymbol{S}) \leqslant x]$. Letting $k \to \infty$ and observing that $\lim_{k \to \infty} \phi^{0,1-k}_j(\boldsymbol{S}) = \phi^0_j(\boldsymbol{S}) < \infty$ almost surely (which implies convergence in distribution too), we get the result (16).

To see why the $(j+1)$st job collides with the $j$th one infinitely often, we argue by contradiction. If this is not true, the two jobs should never collide after some finite index queue $k_0$ (almost surely). Thus, for $k > k_0$, we should have $t_{j+1}^k = t_{j+1}^{k_0} + \sum_{\nu=k_0+1}^{k} \sigma_{j+1}^{\nu}$, so $0 \leqslant t_{j+1}^k - t_j^k = t_{j+1}^{k_0} + \sum_{\nu=k_0+1}^{k} \sigma_{j+1}^{\nu} - t_j^k$ pathwise. Dividing by $k$ and letting $k \to \infty$, we get (see (39) and relevant remarks below)

$$0 \leqslant 0 + E[\sigma_{j+1}^0] - \max_{1 \leqslant i \leqslant j} \{E[\sigma_i^0]\} < 0, \tag{38}$$

which is a contradiction. Thus, the jobs must collide infinitely often almost surely.

To get (38), we use Birkhoff's ergodic theorem [8] to get the term $E[\sigma_{j+1}^0]$, and for the other one we use the formula

$$\lim_{k \to \infty} \frac{t_j^k}{k} = \max_{1 \leqslant i \leqslant j} \{E[\sigma_i^0]\} \tag{39}$$

(almost surely) for every $j \in \mathscr{J}$, which can be proven using (9). To see this intuitively, observe that as $k \to \infty$, the term that gives the dominant contribution in (9) (out of the $j$ terms) is the one which converges to the maximum of all the $E[\sigma_i^0]$'s. For a rigorous proof, see [1], where the result is proven in a related dual context, which is elaborated in Remark 3. This completes the proof of part (a).

(b) To prove (17), observe that analogously to (9) we can write

$$D_a^{n,m} = \max_{1 \leqslant \pi_1 \leqslant \pi_2 \leqslant \pi_3 \leqslant \cdots \leqslant \pi_{n-m} \leqslant a} \left\{ \sum_{p=1}^{\pi_1} \sigma_p^m + \sum_{p=\pi_1}^{\pi_2} \sigma_p^{m+1} + \sum_{p=\pi_2}^{\pi_3} \sigma_p^{m+2} + \cdots + \sum_{p=\pi_{n-m}}^{a} \sigma_p^n \right\} \tag{40}$$

for every $m < n$ and $a \in \mathscr{Z}_+$. Arguing as in (39), we get $\lim_{m \to -\infty} D_a^{n,m}/n - m = \max_{1 \leqslant i \leqslant a} \{E[\sigma_i^0]\}$ almost surely, and because of the assumption in part (b) we get

$$\max_{1 \leqslant i \leqslant j+1} \{E[\sigma_i^0]\} = E[\sigma_{j+1}^0], \tag{41}$$

so

$$\lim_{m \to -\infty} \frac{\phi_j^{n,m}}{n-m} = \lim_{m \to -\infty} \frac{D_{j+1}^{n,m} - D_j^{n,m}}{n-m} = E[\sigma_{j+1}^0] - \max_{1 \leqslant i \leqslant j} \{E[\sigma_i^0]\} > 0 \tag{42}$$

almost surely, which implies that $\phi_j^n = \lim_{m \to -\infty} [\phi_j^{n,m}] = \lim_{m \to -\infty} [D_{j+1}^{n,m} - D_j^{n,m}] = \infty$, proving relation (17).

We can easily get relation (18) by recalling that $\tau_j^k = t_{j+1}^k - t_j^k$ and using (39) (for $(j+1)$ and $j$) and also (41).

Finally, arguing by contradiction, suppose that the $(j+1)$st job collides with the $j$th infinitely often and let $\{k_\rho, \rho \in \mathscr{Z}_+\}$ be the (increasing) sequence of indices of the queues where these two jobs collide. Then, we have $\tau_j^{k_\rho} = \sigma_{j+1}^{k_\rho}$ for every $\rho \in \mathscr{Z}_+$. However, using Birkhoff's ergodic theorem, we get

$$\lim_{\rho \to \infty} \frac{\sigma_{j+1}^{k_\rho}}{k_\rho} = \lim_{\rho \to \infty} \frac{\sum_{\nu=1}^{k_\rho} \sigma_{j+1}^{\nu}}{k_\rho} - \frac{\sum_{\nu=1}^{k_\rho - 1} \sigma_{j+1}^{\nu}}{k_\rho} = E[\sigma_{j+1}^0] - E[\sigma_{j+1}^0] = 0 \tag{43}$$

almost surely, which implies that $\lim_{\rho \to \infty} \tau_j^{k_\rho}/k_\rho = 0$ almost surely, contradicting (18). Therefore, the jobs collide only finitely often. This completes the proof of the theorem. □

It should be noted that in part (a) of the theorem, the jobs colliding infinitely often is *not* enough to guarantee that the interdeparture times asymptotically have proper distributions. Indeed, given the stationary ergodic nature of the service times, it is possible that the jobs collide infinitely often, yet $\lim_{k \to \infty} P[\tau_j^k < \infty] = P[\phi_j^0 < \infty] = 0$. This is exactly what happens in the simple special case where all jobs have the same i.i.d. service times (see Remark 3, last paragraph), due to the null recurrence of the interdeparture times processes in this case. Therefore, to show the finiteness part of the theorem we need to start from first principles, as is done in the proof.

We continue with a few interesting remarks on the previous analysis.

**Remark 1** (*The critical case*). Theorem 1 does not discuss the case where $E[\sigma^0_{j+1}] = \max_{1 \leqslant i \leqslant j}\{E[\sigma^0_i]\}$; we call such a job $j+1 \in \mathcal{J}$ a *critical* one. The proof first collapses at relations (27) and (28), since no contradiction is now encountered. In such a case we cannot say anything about the finiteness of $\phi^n_j$ or whether the $(j+1)$st job collides with the $j$th one infinitely often, based only on the average service times; more information on the distributions is actually needed. As mentioned above, the problem is settled if all service times are i.i.d. (see Remark 3, last paragraph). In the general stationary ergodic case, if there are critical jobs, we can only characterize the noncritical ones by applying the theorem individually on them.

**Remark 2** (*Clustering of jobs*). Interpreting the results of our analysis, we see that, as the jobs move through the queues, they asymptotically form clusters of colliding consecutive jobs; distinct clusters cease to interact eventually. Two consecutive jobs $j$ and $j+1$ belong to the same cluster if $E[\sigma^0_{j+1}] < \max_{1 \leqslant i \leqslant j}\{E[\sigma^0_i]\}$, in which case $\phi^n_j < \infty$ and the two jobs collide infinitely often. If $E[\sigma^0_{j+1}] > \max_{1 \leqslant i \leqslant j}\{E[\sigma^0_i]\}$, then $j+1$ is the first job of a new cluster, $j$ is the last job of the previous one, and $\phi^n_j = \infty$ and the two jobs cease to collide eventually (that is what separates the clusters). Therefore, a cluster is characterized by a front job $j_f$, such that $E[\sigma^0_{j_f}] > \max_{\{0 \leqslant i \leqslant j_b - 1\}}\{E[\sigma^0_i]\}$, and a back (last) job $j_b$, such that $E[\sigma^0_i] < E[\sigma^0_{j_f}]$ for every $j_f < i \leqslant j_b$ and $E[\sigma^0_{j_b+1}] > E[\sigma^0_{j_f}]$ (assuming there are no critical jobs; the critical case is discussed later). All jobs $i \in \{j_f, j_f + 1, \ldots, j_b - 1, j_b\}$ belong to the same cluster.

As the jobs move along the tandem, eventually the clusters become clearly formed and separated, while the distances between the jobs in the same cluster (interdeparture times) reach proper stationary distributions and the jobs collide repeatedly. Actually, all jobs within the same cluster move asymptotically with the same average speed (number of queues crossed per time unit), which we call *cluster speed*. On the contrary, consecutive clusters move at different speeds, the front one always traveling faster than the one following it. Therefore, the relative distance between consecutive clusters grows linearly with time. If job $j$ is left to move along the tandem alone (without interacting with the others), then its (average) speed is $1/E[\sigma^0_j]$, which we call *free speed*. Note that the speed of a cluster is equal to the free speed of its front job, while all the other jobs in it have higher free speeds. This is consistent with intuition about the cluster formation.

A problem arises with respect to critical jobs, defined in Remark 1. We cannot say whether a critical job starts a new cluster or is merged into the one preceding it. An intuitive view is to consider the critical job as loosely coupled to the cluster preceding it, since it does move with the same speed as the cluster, but still it may never collide with it, or its distance from it may grow to infinity asymptotically.

**Remark 3** (*A dual queueing network. The i.i.d. service times case*). An interesting observation is that the dynamics of the infinite queueing tandem can be equivalently described by a *dual* tandem network [4], i.e. one having a finite number of queues and an infinite number of jobs, taking the equivalent point of view of the servers being queued up waiting to access the jobs and keep them for a random holding time. The dual network is constructed as follows. Each job in the original tandem is associated with a FCFS queue (with infinite buffer) in the dual network, and each queue with a token. Therefore, the queues in the dual network are indexed by $j \in \mathcal{J} = \{1, 2, 3, \ldots, J\}$ and the tokens by $k \in \mathcal{Z}_+ = \{1, 2, 3, \ldots\}$, which is the reverse of the original situation. These $J$ queues are connected in tandem, so that a token leaving the $j$th queue joins the $(j+1)$st one. The holding time (staying time) of the $k$th token on the server of the $j$th queue is $\sigma^k_j$. Originally all the tokens are placed in the (infinite) buffer of the first queue ($j = 1$), while all other queues are empty. At time 0 we let the tokens start flowing on the dual network of the $J$ queues. The time when the $k$th token exits the $j$th queue of the dual network is actually the departure time of the $j$th job from the $k$th queue of the original tandem, which is $t^k_j$. Therefore, the quantity $\tau^k_j = t^k_{j+1} - t^k_j$ is actually the delay that the $k$th token will experience in the $(j+1)$st queue while traveling through the dual network.

Observing the output of the first queue of the dual network, we see that the interdeparture times of the tokens are just $\sigma^k_1$, forming a stationary and ergodic arrival process for the rest of the network. Such tandem networks of $G/G/1$ queues have been studied by Loynes [5], extending his stability theory of the G/G/1

queue. If $E[\sigma_i^0] > \max_{1<i\leqslant J}\{E[\sigma_i^0]\}$, the dual network is stable, and Loynes's approach immediately gives the same results as those of Theorem 1. However, if $E[\sigma_{j+1}^0] > \max_{1<i\leqslant j}\{E[\sigma_i^0]\}$ for some $j \in \mathscr{J}'$, then the dual network is unstable and Loynes approach cannot be applied directly to yield the results. In case we have no critical jobs (defined in Remark 1) in the original tandem, a simple extension of Loynes' arguments suffices to produce the results, using the dual network.

Serious problems arise, however, when there are *critical* jobs, making the dual network approach fail and rendering the direct proof presented above necessary. The source of the problems is that in a critically loaded G/G/1 queue the delay process may never merge (pathwise) with a stationary operational regime in finite time (contrary to what happens in subcritical queues) and the output (interdeparture times) may never become stationary (it can actually retain the memory of its initial conditions forever). As a result, when we try to apply Loynes' procedure (which is inductive on the set of queues) to the dual network, we can duplicate the results of the theorem until we reach a critical queue (corresponding to a critical job in the original network). At that point the procedure fails, and nothing can be said about the noncritical queues (jobs) that may follow. This is way we need to take a global approach in the proof of the theorem, and work backwards from $j$, defining the $j^*$. This proof allows us to characterize all the noncritical jobs on the tandem.

If the service times $\sigma_j^k$, $j \in \mathscr{J}$, $k \in \mathscr{Z}$ are mutually independent random variables, we can use the dual network approach to get the results. Indeed, as Lindley [5] and Loynes [6] have proven, a critically loaded queue is always unstable (so in the original infinite tandem the corresponding $\phi_j^n = \infty$ almost surely), but becomes empty infinitely often (so the corresponding jobs collide infinitely often), due to the fact that its workload process is *null recurrent*. The problems discussed above do not appear, since no ambiguity arises, even at critical jobs (the system does not retain memory). Also, using the dual network and Lindlay's and Loynes' results (for the i.i.d. case), we can immediately see that the distances of jobs belonging to the same cluster are positive recurrent, while for different clusters they are transient processes. Such characterizations cannot be established for the general case of stationary ergodic service times, which is the focus of this paper.

An interesting case is when all jobs have independent and identically distributed service times; thus, they are all critical. Then, all distances between the jobs (interdeparture times) are *null-recurrent* processes, and a single cluster of loosely coupled jobs is formed, according to the discussions in Remarks 1 and 2. Using the dual network and Loynes' results, we get that all jobs collide infinitely often; however, all $\phi_j^n$'s are infinite, and $\lim_{k\to\infty} P[\tau_j^k < \infty] = 0$. Glynn and Whitt [4] have further studied the asymptotics of the single cluster formed in that special case, with respect to central limit theorems.

**Remark 5** (*Invariant measures of queueing operators*). From (11) we get $\{\phi_j^{1,m},\ j \in \mathscr{J}'\} = F^1(\{\phi_j^{0,m},\ j \in \mathscr{J}'\}) = F(\{\sigma_j^1,\ j \in \mathscr{J}\},\ \{\phi_j^{0,m},\ j \in \mathscr{J}'\})$, and taking the limits when $m \to -\infty$, we get $\{\phi_j^1,\ j \in \mathscr{J}'\} = F(\{\sigma_j^1, j \in \mathscr{J}\}, \{\phi_j^0, j \in \mathscr{J}'\}) = F^1(\{\phi_j^0, j \in \mathscr{J}'\})$, due to the continuity of the queueing operations involved in $F$. Since $\{\phi_j^1\}$ is identically distributed as $\{\phi_j^0\}$, we see that $\{\phi_j^0\}$ is an invariant measure of the queueing operator $F^1$. From a practical point of view, given the distribution of service times $\{\sigma_j, j \in \mathscr{J}\}$ of a queue, $\{\phi_j^0, j \in \mathscr{J}'\}$ is the distribution of interarrival times that induces an identical distribution of interdeparture times. The existence of such a distribution was proven by Bambos and Walrand [2]. The analysis here can additionally characterize this invariant distribution with respect to finiteness of its components.

**Remark 6** (*Limited buffer space. Blocking*). If each queue has $B$ buffer places, which are less than the total number of jobs $J$, then the possibility of blocking arises, where a job finishing execution at some queue cannot jump to the next one because that is full. The job has to wait until an empty space appears in the following buffer, before it can jump. While waiting, it blocks the server, which resumes service only when the blocked job leaves and the next one enters the service area (like in certain manufacturing systems, where products circulate on pallets). Theorem 1 is still valid here. Specifically, basic relation (13) still holds, and the key arguments (25), (26), (28), (29), (33), (37) are actually amplified by the added restriction of blocking. Instead of using, however, the compact relations (39) and (42), which are valid for infinite buffers, we have to argue (38) and (42) from first principles as in steps (1) and (2) in the theorem; the result is again the same though. Intuitively, we can see that blocking will potentially spread out each cluster, but cannot cause

merging of any two of them. The reason is that the average cluster speed is determined by the free speed of its front job, which eventually ceases to collide with the tail of the preceding cluster, and thus ceases being blocked and attains its free speed.

## 3. Conclusions and final remarks

The asymptotic flow of jobs on an infinite queueing tandem has been studied under stationary and ergodic service times, given that the jobs are simultaneously released at the first queue. It has been shown that stationary clusters are eventually formed of consecutive jobs colliding infinitely often.

Of special interest is a version of the problem where the service times are i.i.d. and the jobs arrive at the first buffer with i.i.d. interarrival times. Again, we want to characterize the asymptotic flow. Although the analysis here provides considerable insight into this modified problem (especially on its sample path structure), serious technical issues remain to be resolved before analogous results can be established. We are working in this direction.

## Acknowledgements

We thank Prof. Thomas M. Liggett and Prof. Roberto Schonmann of the Mathematics Department of UCLA for many insightful discussions concerning the infinite tandem queueing problem. We also thank the referee for a thorough review that has led to the improvement of the paper.

## References

[1] N. Bambos and J. Walrand, On the asymptotic execution time of multi-tasked processes on tandem processors, *Systems Control Lett.* **13** (1989) 391–396.
[2] N. Bambos and J. Walrand, An invariant distribution for the G/G/1 queueing operator, *Adv. in Appl. Probab.* **22** (1990) 254–256.
[3] R. Bellman, A.O. Esogbue and I. Nabeshima, *Mathematical Aspects of Scheduling and Applications* (Pergamon, Oxford, 1982).
[4] P.W. Glynn and W. Whitt, Departures from many queues in series, *Ann. Appl. Probab.* **1** (1991) 546–572.
[5] D.V. Lindley, The theory of queues with a single server, *Proc. Cambridge Philos. Soc.* **48** (1951) 277–289.
[6] R.M. Loynes, The stability of a queue with non-independent interarrival and service times, *Proc. Cambridge Philos. Soc.* **58** (1962) 497–520.
[7] J. Walrand, *An Introduction to Queueing Networks* (Prentice-Hall, Englewood Cliffs, NJ, 1988).
[8] P. Walters, *An Introduction to Ergodic Theory* (Springer, Berlin, 1982).