The Annals of Applied Probability
1994, Vol. 4, No. 3, 693–718

INSTABILITY OF FIFO QUEUEING NETWORKS WITH QUICK SERVICE TIMES[1]

BY MAURY BRAMSON

University of Wisconsin

A class of open first-in, first-out queueing networks is examined. Customers arrive according to a rate-1 Poisson process and wait at queues along their prescribed routes for exponential holding times, after which they exit from the system. Such a network can be chosen so that the sum of the mean service times at each queue is as small as desired. It is shown here that these networks are nevertheless unstable. Each such network will possess two customer types, which proceed along nearly parallel routes. Queues are visited sequentially, with each consisting of one relatively slow step and then several quick steps.

1. Introduction. There has been considerable interest recently within queueing theory in the existence/nonexistence of equilibria for queueing networks. It seems intuitively clear that equilibria will exist for systems whose customers are served substantially more quickly than the rate at which they enter. Despite recent attempts, general results in this direction are lacking. We present here a class of examples which contradict this intuition. Hopefully, they will shed some light on appropriate conditions for the existence of equilibria for general systems of queues.

Consider the general class of open queueing networks with customer types $h = 1, \ldots, H$, where customers enter the system according to independent rate-ν_h Poisson processes, and ν_h is normalized so that $\sum_{h=1}^{H} \nu_h = 1$. Each customer proceeds along a prescribed *route*, visiting a subset of the m queues, $m \geq 1$, and then exiting from the system. Customers are served one at a time at each queue, with the service times being independent and exponentially distributed. The route may depend on the customer type, and individual queues may be visited more than once. The rate a customer is served at a given queue may depend on the position, or *stage*, along the route. Denote by λ_{hij} the rate for the jth visit to the ith queue by a customer of type h. There is also the issue of a priority for the order of service among customers at a given queue. A natural assumption is that the network is *first-in, first-out* (FIFO), that is, customers at a given queue are served in the order they arrive there, irrespective of the customer type or the number of visits previously made.

Received August 1993; revised December 1993.
[1]Supported in part by NSF Grant DMS-93-00612.
AMS 1991 *subject classifications*. 60K25, 68M20.
Key words and phrases. Queueing networks, equilibrium distribution, instability, first-in, first-out.

A fundamental question is under what conditions on λ_{hij} do equilibria exist for the above FIFO queueing networks. If no equilibrium exists, the network is *unstable*. For this, we introduce the notation $\mu_{hij} = \lambda_{hij}^{-1}$, $J(h,i)$ = the number of times the ith queue is visited along the route of a customer of type h, $\mu_{hi} = \sum_{j=1}^{J(h,i)} \mu_{hij}$ and $\mu_i = \sum_{h=1}^{H} \nu_h \mu_{hi}$. Here, μ_{hij} is the mean service time for a customer during a visit at a queue, and μ_i can be interpreted as the total mean service time at i. One can check that if $\mu_i \geq 1$ for some i, then the network is unstable.

A natural conjecture is that an equilibrium will exist under the condition

$$\mu_i < 1 \quad \text{for all } i. \tag{1}$$

It is well known that this is, in fact, true if μ_{hij} does not depend on h and j, and one can, in fact, explicitly write down the equilibrium distribution. [A general theory based on reversibility, which includes FIFO networks, is presented in [3].] However, (1) does not suffice in general, as shown by Bramson [1]. Related work on deterministic systems is given by Seidman [6] and Seidman and Yershov [7], and for priority queues by Lu and Kumar [4] and Rybko and Stolyar [5]. A somewhat more detailed version of the preceding summary is presented in [1].

The above work raises the question of the nature of the *critical value* of the total mean service times μ_i, below which an equilibrium must exist, irrespective of the particular network under consideration. A basic feature of the construction employed in [1] was that $\max_i \mu_i$ needed to be close to 1 to produce an unstable system. Here, we exhibit a class of unstable FIFO networks such that for any given $\mu > 0$, $\mu_i \leq \mu$ for all i is satisfied by an appropriate member of the class.

The networks are assumed to possess m queues, labelled $1, \ldots, m$. Upon entering the system, the two types of customers move along the prescribed routes

$$\begin{aligned} &1 \to 2 \to \cdots \to 2 \to 3 \to \cdots \to 3 \to \cdots \to m \to \cdots \to m \\ &1 \to 1 \to 2 \to \cdots \to 2 \to 3 \to \cdots \to 3 \to \cdots \to m \to \cdots \to m \to 1 \end{aligned} \tag{2}$$

at the end of which they exit from the system. Each portion $i \to \cdots \to i$ of the route consists of seven visits to the ith queue. We refer to customers employing the upper route as *upper* customers, and the others as *lower* customers. Note that the routes for both types of customers are similar, the only difference being that lower customers visit the first queue an extra time both at the beginning and at the end of the route. We denote the stage of a customer by (h,i,j), with the first coordinate being the customer type, the second coordinate the queue and the third coordinate the number of times the queue has been visited up to then. Here $h = u, l$, $i = 1, \ldots, m$ and $j = 1, \ldots, 7$ for $i = 2, \ldots, m$; $j = 1$ for $h = u$ and $i = 1$; and $j = 3$ for $h = l$ and $i = 1$. Each type of customer is assumed to enter the system at rate $\nu_h = \frac{1}{2}$.

The mean service time for a customer is given by

$$
(3)\qquad
\begin{array}{ll}
c \text{ at } (h,i,1), & \text{for } h = u,l \text{ and } i = 2,\dots,m,\\
c \text{ at } (l,1,3), & \\
\delta \text{ at } (h,i,j), & \text{for } h = u,l,\ i = 2,\dots,m \text{ and } j = 2,\dots,7,\\
\delta \text{ at } (u,1,1),(l,1,1) \text{ and } (l,1,2). &
\end{array}
$$

We will assume that

$$
(4)\qquad 0 < c \le 1/100, \qquad 0 \le \delta \le c^8, \qquad m = \left[2c^{-1}\log(c^{-1})\right],
$$

where $[\cdot]$ denotes the integer part. Each of the queues $2,\dots,m$ therefore has one comparatively slow, and six very quick stages for each type of customer. The first queue has only the single quick stage for the upper customers, and two quick stages and one slow stage for lower customers. The choice of parameters is made for technical reasons. [Each portion $i \to \cdots \to i$ of the route in (2) can be reduced to four visits, but then the upper bound in (4) for c would have to be much smaller for our estimates to work and the accompanying mechanics more complicated; the bound $c \le 1/100$ is somewhat arbitrary. The coefficient 2 in the definition of m can be replaced by any value in $(1,4)$. In our case, $(1-c)^{-m} \sim c^{-2}$, which is used in Step 3 at the end of Section 2. Note that if $m < (c+6\delta)^{-1}$, then $\sum_i \mu_i < 1$, in which case the system in (2)–(3) must have an equilibrium.]

Under (4), $\mu_i \le c + 6\delta < 2c$, so the total mean service time can be chosen as small as desired at every i. Our main result is that the following theorem nevertheless holds.

THEOREM 1. *Any FIFO queueing network of the form* (2), *and satisfying* (3) *and* (4), *is unstable, with the number of customers in the system approaching infinity as* $t \to \infty$.

Theorem 1 has the following consequence. One can compare any FIFO queueing network satisfying (2)–(4) with the network which is obtained from it by replacing (3) with the simple assumption that the mean service time $\mu_{hij} \equiv c$ at every stage of the route. The lengths of the mean service times for the new network are, of course, everywhere at least as great as for the original network. As mentioned earlier, this new queueing network has an equilibrium distribution which can be written down explicitly; see, for example, page 61 of [3]. (One can check that the equilibrium probability of k customers at any given queue is at most $(1-7c)(7c)^k$, $k \ge 1$, which means that the network is in fact "very stable" for small c.) This observation shows that decreasing mean service times within a queueing network may have the effect of making it unstable.

Theorem 1 and the above comparison provide an explanation for the lack of general criteria so far on queueing networks which ensure the existence of equilibria. It is perhaps risky to propose criteria at this point, but the

following formulations suggest themselves. We set $\mu_i^{\min}(\mu_i^{\max})$ equal to the minimum (maximum) over all μ_{hij}, where i is fixed, and $\mu_i^R = \mu_i^{\min}/\mu_i^{\max}$.

QUESTION 1. For each given $\mu^R > 0$, does there exist a $\mu > 0$, so that any FIFO network satisfying $\mu_i^R \geq \mu^R$ and $\mu_i \leq \mu$ for all i has an equilibrium?

QUESTION 2. For each given $\mu < 1$, does there exist a $\mu^R < 1$, so that any FIFO network satisfying $\mu_i^R \geq \mu^R$ and $\mu_i \leq \mu$ for all i has an equilibrium?

An affirmative answer to Question 1 would say that if the mean service times at a given queue are not too different, then small enough values of μ_i suffice for an equilibrium. An affirmative answer to Question 2 would say that the system has an equilibrium for $\mu_i < 1$ as long as μ_i^R are close enough to 1. The networks with μ_{hij} not dependent on h and j satisfy $\mu_i^R \equiv 1$, and so make up the limiting case in the latter setup. (By arguing as in [2], it is not difficult to show that the behavior addressed in Question 2 holds for FIFO networks of fixed length.) One can, of course, restrict these questions to networks with only one type of customer. In this context, the situation is also unclear. (Theorem 1 can presumably be replaced by a single customer type analog, from which similar conclusions would follow.)

The remainder of the paper is structured as follows. In Section 2, we present a summary of the induction argument upon which Theorem 1 is based, and show how Theorem 1 itself follows. Two key estimates for the main recursion argument are derived in Section 3. The recursion argument itself is treated in Section 4. In Section 5, it is shown that under appropriate initial conditions, the evolution is nearly periodic, with the network returning to an amplified version of its original state. Repetition of this procedure demonstrates the instability of the network.

2. Summary of the proof. We first introduce some notation. Let Ξ_t denote the state at time t of a queueing network satisfying (2)–(4), and let $\xi_t(h,i,j)$ denote the number of customers at (h,i,j) at time t. We employ (i,j) for the union of (u,i,j) and (l,i,j). We set $\xi_t(i,j) = \xi_t(u,i,j) + \xi_t(l,i,j)$ and $\xi_t(i) = \sum_j \xi_t(i,j)$. [Here, $\xi_t(1,j) = \xi_t(l,1,j)$, for $j = 2,3$.] We denote by ξ_t the total number of customers in the system. By $(i,j)^+$ [resp., $(i,j)^-$], we will mean the set of stages in the system strictly beyond [resp., before] (i,j), and by $\xi_t(i,j)^+$ [resp., $\xi_t(i,j)^-$], the number of customers in $(i,j)^+$ [resp., $(i,j)^-$]. [Here $(i,j)^+$ and $(i,j)^-$ do not distinguish between the upper and lower routes.] For instance,

$$\xi_t(i,1)^- = \sum_{j=1}^{2} \xi_t(1,j) + \sum_{i'=2}^{i-1} \xi_t(i'), \qquad i = 2,\ldots,m.$$

The proof of Theorem 1 is based on the following induction step. Here and later on, we set $\xi_0(u,1,1) = M_u$, $\xi_0(l,1,1) = M_l$ and $M = M_u + M_l$.

THEOREM 2. *Assume that*

$$(5)\quad \tfrac{1}{4}M \le M_u \le \tfrac{3}{4}M, \qquad \xi_0(1,1)^+ \le c^4 M, \qquad \xi_0(1,2) + \xi_0(2) \le c^5 M.$$

Then for some $\varepsilon_1 > 0$, *large enough* M *and appropriate random* T (*depending on* M),

$$(6)\quad \begin{aligned} P\Big[&\tfrac{1}{4}\xi_T(1,1) \le \xi_T(u,1,1) \le \tfrac{3}{4}\xi_T(1,1), \\ &\xi_T(1,1) \ge \tfrac{1}{4}c^{-1}M,\ \xi_T(1,1)^+ \le \tfrac{1}{8}c^3 M, \\ &\xi_T(1,2) + \xi_T(2) \le \tfrac{1}{8}c^4 M\Big] \ge 1 - \exp(-\varepsilon_1 M) \end{aligned}$$

and

$$(7)\quad P\big(\xi_t \ge \tfrac{1}{5}M,\ \forall\, t \in [0,T]\big) \ge 1 - \exp(-\varepsilon_1 M).$$

We will later choose T to be about $c^{-2}M$. One has some leeway in the choice of bounds for $\xi_0(1,1)^+$ and $\xi_0(1,2) + \xi_0(2)$. If the latter is of the form $c^\alpha M$, with $\alpha \in [5,6)$, and the former $c^{\alpha'}M$, with $\alpha' \in (1,\alpha)$, then bounds similar to (6) hold.

One can show that Theorem 1 follows from Theorem 2. Suppose that Ξ_0 satisfies (5) for some large M. Since $c \le 1/100$, one has $\frac{1}{4}c^{-1}M \ge 25M$. Moreover, the stated bounds for $\xi_t(1,1)$ relative to $\xi_t(1,1)^+$ and $\xi_t(1,2) + \xi_t(2)$ improve from (5) to (6). So, repeated application of Theorem 2 yields

$$(8)\quad P\big(\xi_t < \tfrac{1}{5}M \text{ for some } t \ge 0\big) \le 2\sum_{k=0}^{\infty} \exp\Big[-(25)^k \varepsilon_1 M\Big],$$

which approaches 0 as $M \to \infty$. Since all states in the system are accessible from one another, (8) implies that $\xi_t \to \infty$ as $t \to \infty$ w.p.1 for any Ξ_0. Theorem 1 then follows.

We have up to now followed the format given in [1]. The queueing networks considered there had one customer type, with prescribed route $1 \to 2 \to 2 \to \cdots \to 2 \to 1$, the number of visits to the second queue being large. The mean service times were given by c at $(1,2)$ and $(2,1)$, and δ at other stages, where c was close to 1 and δ was very small. (The notation corresponds to that introduced in Section 1.) Since our present system is by nature more complicated, with a large number of queues, different reasoning must be employed here to demonstrate Theorem 2. The basic building block is, however, the same in both cases, with the concept of "cycles" being used to regulate the movement of customers throughout the system.

Customers at any given queue at time t may be ordered according to the times at which they are next served, so one can talk about a "first" or "last" customer in this sense. (Due to the multiple stages at each queue, customers entering the network earlier on may be ordered behind more recent arrivals. Upper and lower customers are ordered without distinction.) Let $S_{1,1}$ denote the time at which the last of the original customers (customers at $t = 0$) at

the first queue is next served. Similarly, let $S_{2,1}$ denote the time when the last of the customers at the first queue is served, with the ordering this time being made at $t = S_{1,1}$. [By $S_{2,1}$, all of the customers originally at $(1,1)$ and $(1,2)$ have moved to at least the second queue.] Denote by $S_{2,2}, \ldots, S_{2,7}$ the times at which the last customer at the second queue is served, with the orderings being made at $t = S_{2,1}, \ldots, S_{2,6}$. Let $S_{3,1}, S_{3,1} \geq S_{2,7}$, denote the next time at which the second queue is empty. [On account of (1), $S_{3,1} < \infty$ w.p.1.] Proceeding inductively, one denotes by $S_{i,2}, S_{i,3}, \ldots, S_{i,7}$ the times at which the last customer at the ith queue is served, with the orderings being made at $t = S_{i,1}, \ldots, S_{i,6}$, and by $S_{i+1,1}$ the next time at which the ith queue is empty. In this manner, one defines random times up through $S_{m,2}, \ldots, S_{m,7}$ and $S_{1,2}$, where $S_{1,2} \geq S_{m,7}$ is the next time at which the mth queue is empty. Last, let T (as in Theorem 2) denote the time when the last of the customers at the first queue is served, the ordering being made at $t = S_{1,2}$. Note that $S_{1,1}, \ldots, T$ are all stopping times for Ξ_t. We can think of the intervals $(S_{i,j}, S_{i,j+1}]$, $j = 1, \ldots, 6$, as *cycles*, over which each customer starting in the ith queue is served exactly once by this queue. Together with $(S_{i,7}, S_{i+1,1}]$, at the end of which the ith queue is empty, these intervals regulate the movement of customers. In particular, as we will see, most of the customers in the system at $t = S_{i,j}$, $i \geq 2$, are at (i,j). (The interval $(S_{i,7}, S_{i+1,1}]$ will be too short for any but a few of the customers to move beyond $(i+1, 1)$ by $t = S_{i+1,1}$.) Also, at $S_{1,1}$, most of the customers are at $(u, 2, 1)$ and $(l, 1, 2)$, at $S_{1,2}$, most are at $(l, 1, 3)$ and at T, most are at $(1, 1)$.

The presence of large numbers of customers at $(1, 1)$ at time T, as asserted in (6), is induced by the large numbers of customers at the end of the lower route, $(l, 1, 3)$, at $t = S_{1,2}$, and by the time required to serve them. The purpose of including upper customers in the network is to restrict the movement of the lower customers. The regulated movement of customers over the successive cycles just enumerated will allow, as we will show, only comparatively few upper customers to advance more quickly than the main body of customers. Because of the absence of a $(u, 1, 2)$ stage, lower customers entering the system at about the same time as upper customers will lag behind the upper customers as both move through the system. No lower customers are therefore able to advance quickly through the system. In particular, none reaches the stage $(l, 1, 3)$ "ahead of schedule." Since the movement of customers entering the system is delayed only by this final stage, the presence of upper customers prevents a feedback effect which could conceivably throw the evolution of the system out of control.

We utilize these observations to summarize the argument for Theorem 2. The following five steps illustrate the main ideas, although the reader should keep in mind they involve some oversimplifications. In particular, we are ignoring the contributions of the exceptional events over which the individual steps fail to hold. (They are exponentially small in M.)

1. $S_{1,1}$ and $S_{2,1}$ are small relative to M. At $t = S_{2,1}$, most customers in the system are at $(2, 1)$.

2. As previously stated, most customers in the system move from $(2,1)$ to $(3,1)$ over $(S_{2,1}, S_{3,1}]$. The number of these customers increases from about M to about $(1-c)^{-1}M$. Hence $S_{3,1}$ is about $c(1-c)^{-1}M$.
3. Arguing by induction, one obtains that the number of customers moving from $(i-1,1)$ to $(i,1)$ over $(S_{i-1,1}, S_{i,1}]$ is about $(1-c)^{2-i}M$, and $S_{i,1} - S_{i-1,1}$ is about $c(1-c)^{2-i}M$. Summing over these times, one sees that $S_{1,2}$ is about $\sum_{i=3}^{m+1} c(1-c)^{2-i}M \sim (1-c)^{1-m}M \sim c^{-2}M$. At this time, there are about $\frac{1}{2}c^{-2}M$ customers at $(l,1,3)$; an equal number have already exited from the upper route of the system. There are never enough customers behind or ahead of the main body of customers to affect the computations. (The actual ordering of the reasoning is somewhat different in the proof.)
4. The additional time $T - S_{1,2}$ required for the customers to leave $(l,1,3)$ is about $\frac{1}{2}c^{-1}M$. About $\frac{1}{2}c^{-1}M$ new customers arrive at $(1,1)$ during $(S_{1,2}, T]$ and are equally split between upper and lower types. They remain at $(1,1)$ until the customers at $(l,1,3)$ leave. There are few customers elsewhere in the system. This shows (6) of Theorem 2.
5. Lower customers starting at $(1,1)$ do not leave the system before $t = S_{m,7}$, at which time there are about $\frac{1}{2}c^{-2}M$ lower customers in the system. Since $S_{1,2} - S_{m,7}$ is comparatively small, few customers leave during $(S_{m,7}, S_{1,2}]$. The mechanism in Step 4 therefore guarantees that the number of customers in the system will remain above $\frac{1}{2}c^{-1}M$ until $t = T$. In particular, the number of customers in the system never drops much below $\frac{1}{4}M$ over $[0,T]$. This shows (7) of Theorem 2.

3. Some upper bounds for Ξ_t. We provide part of the machinery here which we will need in order to establish the unstable behavior of Ξ_t outlined in the previous section. The main results are Propositions 1 and 2, which give upper bounds on the growth of the number of customers over the subsets $(i,1)^-$, $i = 2,\ldots,m$, of the network as t increases. These bounds will be employed in Section 4 to show that off of a set of exponentially small probability, there are never enough customers behind the main body of customers to affect the qualitative behavior of Ξ_t. We also analyze the behavior of Ξ_t at $t = S_{2,1}$. The corresponding estimates are not difficult and are based on upper bounds for $S_{1,1}$ and $S_{2,1}$.

We will repeatedly be making use of elementary large deviation estimates for times such as $S_{i,j+1} - S_{i,j}$, $j = 1,\ldots,6$, and $S_{i+1,1} - S_{i,7}$, and for the numbers of customers who have entered and left different stages of the route over these times. These estimates all reduce to applying the strong Markov property in conjunction with the following basic bounds: Let $X_1, X_2, \ldots$ be i.i.d. mean-1 exponential random variables, with $Y_n = X_1 + \cdots + X_n$. Then for each $\alpha > 0$, there exists an $\varepsilon > 0$, so that for all $n \geq 1$,

$$P\left(\frac{1}{n}|Y_n - n| > \alpha\right) \leq \exp(-\varepsilon n). \tag{9}$$

The bound (9) can be demonstrated in the usual way be applying Markov's inequality to the Laplace transform of Y_n. Note that (9) immediately extends to exponential distributions with other means by rescaling. We will be applying (9) throughout the paper, with different choices of α and ε. Rather than do precise bookkeeping with the different values of ε, we will label them as $\varepsilon_1, \varepsilon_2, \ldots$, without worrying about their exact relationship. Here and elsewhere in the paper, our choices of c and δ will be regarded as fixed, and we will not be concerned with their precise effect on such exponents. By "typically," we will mean off sets of exponentially small probability.

Proposition 1 gives us control over the growth of $\xi_t(2,1)^-$ and Proposition 2 gives us control over $\xi_t(i,1)^-$, $i = 3, \ldots, m$. We employ the notation

$$E_i(N;\sigma,\tau) = \left\{\xi_t(i,1)^- \leq N \text{ for all } t \in [\sigma,\tau]\right\}$$

for $i = 2, \ldots, m$, where σ and τ are random times. It will later be convenient to use the normalization

$$F_i(N;\sigma,\tau) = E_i\left(\left(\frac{1+c^2}{1-c}\right)^{i-2} N;\sigma,\tau\right). \tag{10}$$

Note that $F_2(N;\sigma,\tau) = E_2(N;\sigma,\tau)$. We abbreviate $E_i(N;0,\tau)$ by $E_i(N;\tau)$ and $F_i(N;0,\tau)$ by $F_i(N;\tau)$. Here and elsewhere, $(\cdot)^c$ means complement. An important ingredient in Proposition 1 is $\zeta(t)$, the total number of lower customers visiting the stage $(1,3)$ by time $t-$. The proposition states that if neither $\xi_0(2,1)^-$ nor $\zeta(\tau)$ is too large, then $\xi_t(2,1)^-$ will quickly become and then remain small over a long interval of time.

PROPOSITION 1. *Assume that* $\xi_0(2,1)^- \leq N_1$. *For appropriate* $\varepsilon_2 > 0$,

$$P\left(E_2^c(2cN_2; 3\delta N_1 + 2cN_2, \tau);\ \zeta(\tau) \leq N_2\right) \leq t_0 \exp(-\varepsilon N_2)$$

for large enough N_2, $t_0 \geq 0$, *any* N_1, *and any random time* τ *with* $\tau \leq t_0$.

PROOF. The quantity $\xi_t(2,1)^-$ is bounded above by

$$U_t = \xi_t(u,1,1) + 2\xi_t(l,1,1) + \xi_t(l,1,2). \tag{11}$$

Each time a customer enters the system, U_t increases by 1 or 2, and each time a customer in $(2,1)^-$ is served, U_t decreases by exactly 1. We find it convenient to analyze U_t over a modified time scale. Partition $[0,\infty)$ into the two sets I_1 and I_2, where $t \in I_1$ if either the customer being served at the first queue is in $(2,1)^-$ or the queue is empty, and $t \in I_2$ if the customer is at $(l,1,3)$. Write

$$\varphi(t) = |[0,t] \cap I_1|,$$

where $|\cdot|$ denotes Lebesgue measure, and set

$$U'_s = U_{\varphi^{-1}(s)}, \qquad s \geq 0, \tag{12}$$

with $\varphi^{-1}(s)$ chosen to be right continuous. That is, U'_s is the process obtained by ignoring the time over which customers at $(l,1,3)$ are served. The process

U_s' decreases by 1 at rate δ^{-1} if $U_s' > 0$, increases by 1 and 2, each at rate $\frac{1}{2}$, and has jumps given by the number of upper customers plus twice the number of lower customers entering the system over intervals in I_2. Denote by $n(t)$ the weighted sum for customers entering the system over $[0, t] \cap I_2$ (with lower customers counting double).

The process U_s' can be compared with the process V_s, $V_s \geq 0$, which decreases by 1 at rate δ^{-1} if $V_s > 0$ and increases by 2 at rate 1. Assume that $V_0 = U_0' = U_0$. If U_s' and V_s are coupled together so that increases and decreases occur together, then one can check that

$$U_s' \leq V_s + n\big(\varphi^{-1}(s)\big). \tag{13}$$

(Note that the inequality is not only due to the difference in the size of the increases, but also the possible decreases in U_s' when $V_s = 0$.)

Since V_s is a random walk on $\{0, 1, 2, \ldots\}$ with negative drift $2 - \delta^{-1}$ and $V_0 \leq 2N_1$, the following bound follows by standard large deviation techniques: For appropriate $\varepsilon_3 > 0$,

$$P\big(V_s > (c/8)/N_2 \text{ for some } s \in \big[3\delta N_1 + \tfrac{3}{4}cN_2, t_0\big]\big) \leq t_0 \exp(-\varepsilon_3 N_2) \tag{14}$$

for large enough N_2, and any N_1 and t_0. (The factor $\frac{3}{4}c$ is not optimal and is chosen for convenience later on.) One can, for instance, first compare V_s with the corresponding random walk on $\mathbb{Z}$ to show that the time ρ_1 at which V_s hits 0 is at most $3\delta(N_1 + N_2) \leq 3\delta N_1 + \frac{3}{4}cN_2$ off of a set G of probability $\exp(-\varepsilon_5 N_2)$. One can then employ the supermartingale

$$W_{s'} = \exp\{V_{s'+\rho_1}\} - e^2 s'$$

off of G. (The term $e^2 s'$ is needed when $V_{s'+\rho_1} = 0$.) Set

$$\rho_2 = \inf\{s' : V_{s'+\rho_1} \geq (c/8)N_2\} \wedge t_0.$$

By Chebyshev's inequality, the left side of (14) is at most

$$\begin{aligned} &\exp(-cN_2/8)E\big[\exp\{V_{\rho_1+\rho_2}\}; G^c\big] + P(G) \\ &\quad \leq \exp(-cN_2/8)E\big[W_{\rho_2} + e^2\rho_2; G^c\big] + \exp(-\varepsilon_5 N_2). \end{aligned}$$

By the Optional Sampling Theorem, this is at most

$$\begin{aligned} &\exp(-cN_2/8)E\big[W_0 + e^2 t_0; G^c\big] + \exp(-\varepsilon_5 N_2) \\ &\quad \leq \big(1 + e^2 t_0\big)\exp(-cN_2/8) + \exp(-\varepsilon_5 N_2). \end{aligned}$$

This implies (14), since $t_0 \geq \frac{3}{4}cN_2$ can be assumed.

On account of (9), the sum of the service times for the first N_2 customers at $(l, 1, 3)$ is at most $\frac{9}{8}cN_2$ off of a set of exponentially small probability. Since $t - \varphi(t)$ is increasing, it follows that for appropriate $\varepsilon_4 > 0$ and any random time τ,

$$P\big(t - \varphi(t) > \tfrac{9}{8}cN_2 \text{ for some } t \leq \tau; \zeta(\tau) \leq N_2\big) \leq \exp(-\varepsilon_4 N_2). \tag{15}$$

Another application of (9) implies that the numbers of upper and lower customers entering the system over these N_2 service times are each at most $\frac{5}{8}cN_2$, again off a set of exponentially small probability. On $\{\zeta(\tau) \leq N_2\}$, the weighted sum of such customers (with lower customers counting double) is at least $n(\tau)$, and so by (13),

$$(16) \quad P\big(U_s' > V_s + \tfrac{15}{8}cN_2 \text{ for some } s \leq \varphi(\tau);\ \zeta(\tau) \leq N_2\big) \leq \exp(-\varepsilon_5 N_2),$$

for appropriate $\varepsilon_5 > 0$, and large N_2.

Together, (14) and (16) show that

$$(17) \quad \begin{aligned} &P\big(U_s' > 2cN_2 \text{ for some } s \in \big[3\delta N_1 + \tfrac{3}{4}cN_2, \varphi(\tau) \wedge t_0\big];\ \zeta(\tau) \leq N_2\big) \\ &\qquad \leq t_0 \exp(-\varepsilon_6 N_2), \end{aligned}$$

for $\varepsilon_6 > 0$. Always, $\varphi(t) \leq t$, and by (15), $\varphi(t) \geq t - \frac{9}{8}cN_2$ is typically the case for $t \leq \tau$. From (12) and (17), it follows that

$$P\big(U_t > 2cN_2 \text{ for some } t \in \big[3\delta N_1 + 2cN_2, \tau\big];\ \zeta(\tau) \leq N_2\big) \leq t_0 \exp(-\varepsilon_2 N_2),$$

with appropriate $\varepsilon_2 > 0$, for $\tau \leq t_0$. This implies the proposition. □

Assuming that we can control the growth of $\xi_t(i-1,1)^-$ over t, we can also control the growth of $\xi_t(i,1)^-$. This is the content of Proposition 2.

PROPOSITION 2. *Assume that* $\xi_0(i-1) = 0$ *for some* i, $i = 3, \ldots, m$. *For appropriate* $\varepsilon_7 > 0$,

$$P\left(E_{i-1}(N;\tau) \cap E_i^c\left(\left(\frac{1 + \frac{1}{2}c^2}{1-c}\right)(N+x);\tau\right)\right) \leq t_0 \exp(-\varepsilon_7 x)$$

for any x, N, $t_0 \geq 2$, *and any random time* τ *with* $\tau \leq t_0$.

Setting $x = \frac{1}{4}c^2 N$ in Proposition 2 and employing the notation in (10), we obtain the following variant.

COROLLARY. *Assume that* $\xi_0(i-1) = 0$ *for some* i, $i = 3, \ldots, m$. *For appropriate* $\varepsilon_8 > 0$,

$$P\big(F_{i-1}(N;\tau) \cap F_i^c(N;\tau)\big) \leq t_0 \exp(-\varepsilon_8 N)$$

for any N, $t_0 \geq 2$, *and any random time* τ *with* $\tau \leq t_0$.

Repeated application of the corollary, in conjunction with Proposition 1, will show that $\xi_t(i,1)^-$ can only grow more or less like $(1-c)^{-(i-2)}$ under suitable initial conditions. We apply this in Lemma 3 of Section 4.

PROOF OF PROPOSITION 2. We set

$$(18) \quad \begin{aligned} U_t = {} & \xi_t(i-1,2)^- + \big(1 - c(1+c^2)\big)\xi_t(i-1,2) \\ & + \big(1 - (c+\delta)(1+c^2)\big)\xi_t(i-1,3) \\ & + \cdots + \big(1 - (c+5\delta)(1+c^2)\big)\xi_t(i-1,7) - N. \end{aligned}$$

(This U_t should not be confused with the definition given in (11).) Each time a customer enters the system, U_t increases by 1. Service of a customer in the $(i-1)$st queue decreases U_t by at least $\delta(1+c^2)$; service of customers in other queues does not directly affect U_t. For $U_t > 0$, $\xi_t(i,1)^- > N$. Also, set

$$\rho = \inf\{t\colon \xi_t(i-1,1)^- > N\}$$

and

$$\begin{aligned} U_t' &= U_{t\wedge\rho} \quad \text{if } \xi_0(i-1,1)^- \le N, \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

We introduce the process

$$W_t = \exp\bigl(c^2U_t'\bigr) - \tfrac{1}{2}t.$$

The point of the preceding terminology is that W_t is a supermartingale. To check this requires a little bookkeeping. If $U_t' \le 0$ and $t < \rho$, then W_t increases at rate 1 by the amount

$$\exp\bigl(c^2(U_t'+1)\bigr) - \exp\bigl(c^2U_t'\bigr) = \exp\bigl(c^2U_t'\bigr)\bigl(\exp(c^2) - 1\bigr) \le c^2(1+c^2) < \tfrac{1}{2}$$

(since c is small), which is dominated by the contribution from the term $-\frac{1}{2}t$. If $U_t' > 0$ and $t < \rho$, then the $(i-1)$st queue is not empty. Recall that customers at $(i-1,1)$ are served at rate c^{-1}, and those at $(i-1,j)$, $j = 2,\dots,7$, at rate δ^{-1}. The process U_t was chosen in (18) so that irrespective of the stage of the customer being served in this queue, W_t will decrease at rate (weighted for the jump size) at least

$$c^2(1+c^2)(1-c^3)\exp\bigl(c^2U_t'\bigr) \ge c^2\bigl(1+\tfrac{3}{4}c^2\bigr)\exp\bigl(c^2U_t'\bigr). \tag{19}$$

The right side of (19) is an upper bound for the mean rate of increase of W_t due to customers entering the system. So, W_t is a supermartingale. Note that $U_0' \le 0$, and so $E[W_0] \le 1$.

We now set

$$\rho_x = \inf\{t\colon U_t' \ge x\} \wedge t_0, \qquad t_0 \ge 0.$$

For any random time τ, $\tau \le t_0$,

$$P\bigl(U_t \ge x \text{ for some } t \le \tau;\, E_{i-1}(N;\tau)\bigr) \le P\bigl(U_{\rho_x}' \ge x\bigr).$$

By Chebyshev's inequality and the Optional Sampling Theorem, this is

$$\begin{aligned} &\le \exp(-c^2x)\,E\bigl[\exp\{c^2U_{\rho_x}'\}\bigr] = \exp(-c^2x)\,E\bigl[W_{\rho_x} + \tfrac{1}{2}\rho_x\bigr] \\ &\le \exp(-c^2x)\,E\bigl[W_0 + \tfrac{1}{2}t_0\bigr] \le \bigl(1+\tfrac{1}{2}t_0\bigr)\exp(-c^2x). \end{aligned}$$

It follows that for $t_0 \ge 2$,

$$P\bigl(U_t \ge x \text{ for some } t \le \tau;\, E_{i-1}(N;\tau)\bigr) \le t_0\exp(-c^2x). \tag{20}$$

On account of (18),

$$\xi_t(i,1)^- \le (U_t + N)/\bigl(1-(c+5\delta)(1+c^2)\bigr),$$

the denominator being the smallest of the coefficients of the terms $\xi_t(i-1,j)$, $j = 2, \ldots, 7$. Since c is small,

$$\left(1 - (c + 5\delta)(1 + c^2)\right)^{-1} \leq \left(1 + \tfrac{1}{2}c^2\right)/(1 - c).$$

Plugging these bounds into (20) implies the proposition. □

In Sections 4 and 5, we will examine the evolution of Ξ_t over $[0, T]$ under the assumption (5) on Ξ_0. The main steps are given by the induction argument in Section 4. Here, we provide bounds at the time $S_{2,1}$ at which the argument begins. We will find it convenient to use the following terminology. Let $\beta_t(h,i,j)$ [resp., $\gamma_t(h,i,j)$] be the number of arrivals at (resp., departures from) (h,i,j) over $(0,t]$, with $h = u$ and $h = l$. Let $\beta_t(i,j)$ [resp., $\gamma_t(i,j)$] be the sum of both types. Note that $\beta_t(1,1)$ is the number of customers entering the network and $\gamma_t(l,1,3)$ is the number of lower customers leaving the network. We also let $\gamma_t(1,3)$ denote the total number of customers (both upper and lower) leaving the network.

LEMMA 1. *Suppose that* (5) *holds. For appropriate* $\varepsilon_k > 0$, *and large enough* M,

$$P\left(S_{2,1} > 4c^5M\right) \leq \exp(-\varepsilon_9 M), \tag{21}$$

$$P\left(\beta_{S_{2,1}}(1,1) > 6c^5M\right) \leq \exp(-\varepsilon_{10}M) \tag{22}$$

and

$$P\left(\gamma_{S_{2,1}}(2,1) > 5c^4M\right) \leq \exp(-\varepsilon_{11}M). \tag{23}$$

PROOF. We first investigate the behavior at $S_{1,1}$. Customers at $(1,1)$ and $(1,2)$ are served at rate δ^{-1}, and those at $(1,3)$ are served at rate c^{-1}. Applying (5) together with (9) twice (to mean-δ and mean-c exponentials), it follows that for any $\alpha > 0$,

$$P\left(S_{1,1} > \left(\delta(1 + c^5) + c^5 + \alpha\right)M\right) \leq \exp(-\varepsilon_{12}M) \tag{24}$$

for appropriate $\varepsilon_{12} > 0$ and large enough M. Since $\delta(1 + c^5) \leq \frac{1}{2}c^5$, choosing $\alpha = \frac{1}{2}c^5$, one has

$$P\left(S_{1,1} > 2c^5M\right) \leq \exp(-\varepsilon_{12}M). \tag{25}$$

Another application of (9) implies that

$$P\left(\beta_{2c^5M}(1,1) > 3c^5M\right) \leq \exp(-\varepsilon_{13}M), \qquad \varepsilon_{13} > 0.$$

Together with (25), this shows that

$$P\left(\beta_{S_{1,1}}(1,1) > 3c^5M\right) \leq \exp(-\varepsilon_{14}M), \qquad \varepsilon_{14} > 0. \tag{26}$$

By time $S_{1,1}$, the M_u customers originally at $(u,1,1)$ have moved to at least the second queue, and the M_l customers at $(l,1,1)$ to $(l,1,2)$, where

they remain. An upper bound on the number of new customers at $(1,1)$ is given by (26). The same reasoning as in (24) and (25) gives

$$P\big(S_{2,1} - S_{1,1} > 2c^5M\big) \le \exp(-\varepsilon_{15}M) \qquad \text{for } \varepsilon_{15} > 0. \tag{27}$$

Together with (25), this implies (21). The same reasoning as in (26) gives

$$P\big(\beta_{S_{2,1}}(1,1) - \beta_{S_{1,1}}(1,1) > 3c^5M\big) \le \exp(-\varepsilon_{16}M) \qquad \text{for } \varepsilon_{16} > 0.$$

Together with (26), this implies (22). Customers are served at $(2,1)$ at rate c^{-1}. So (23) follows from (21) and (9). □

Using Lemma 1, we can analyze $\xi_{S_{2,1}}(2,1)^-$, $\xi_{S_{2,1}}(2,1)$ and $\xi_{S_{2,1}}(2,1)^+ + \gamma_{S_{2,1}}(1,3)$.

PROPOSITION 3. *Suppose that* (5) *holds. For* ε_{10} *and* ε_{11} *as above, appropriate* $\varepsilon_{17} > 0$ *and large enough* M,

$$P\big(\xi_{S_{2,1}}(2,1)^- > 6c^5M\big) \le \exp(-\varepsilon_{10}M), \tag{28}$$

$$P\big(|\xi_{S_{2,1}}(2,1) - M| > cM\big) \le \exp(-\varepsilon_{17}M) \tag{29}$$

and

$$P\big(\xi_{S_{2,1}}(2,1)^+ + \gamma_{S_{2,1}}(1,3) > 11c^4M\big) \le \exp(-\varepsilon_{11}M). \tag{30}$$

PROOF. By time $S_{2,1}$, all of the customers originally at $(1,1) \cup (1,2)$ are no longer there. Therefore,

$$\xi_{S_{2,1}}(2,1)^- \le \beta_{S_{2,1}}(1,1),$$

and so (28) follows from (22). Note that

$$\xi_{S_{2,1}}(2,1) \ge \beta_{S_{2,1}}(2,1) - \gamma_{S_{2,1}}(2,1).$$

All of the M customers originally at $(1,1)$ have arrived at $(2,1)$ by time $S_{2,1}$, and so $\beta_{S_{2,1}}(2,1) \ge M$. Together with (23), this implies the lower bound in (29). It is easy to check that

$$\xi_{S_{2,1}}(2,1) \le \xi_0(3,1)^- + \beta_{S_{2,1}}(1,1),$$

which by (5) and (22) is at most $(1 + 7c^5)M$ off of the exceptional set. So the upper bound in (29) follows as well. Since

$$\xi_{S_{2,1}}(2,1)^+ + \gamma_{S_{2,1}}(1,3) = \xi_0(2,1)^+ + \gamma_{S_{2,1}}(2,1), \tag{31}$$

(30) follows from (23). (The coefficient 11 is chosen for compatibility later on.) □

4. The induction step. In this section, we track the behavior of Ξ_t over the time interval $(S_{2,1}, S_{m,7}]$, which comprises the major part of $[0,T]$. As we show below, nearly all customers in the network at $t = S_{i,j}$, $i = 2,3,\ldots,m$, and $j = 1,\ldots,7$, will be at (i,j). This explicit structure will enable us to

argue inductively on i and j in Propositions 4 and 5, the main results here. Proposition 4 treats the evolution of Ξ_t over the intervals $(S_{i,1}, S_{i,7}]$, $i = 2, \ldots, m$, and Proposition 5, the evolution over $(S_{i,7}, S_{i+1,1}]$, $i = 2, \ldots, m-1$. Together, they imply the behavior at $t = S_{m,7}$ given in (39)–(41). The rest of the section is devoted to demonstrating the two propositions.

Proposition 4 gives careful bounds on the state of the network at times $S_{i,7}$, $i = 2, \ldots, m$. The actual proof employs analogous bounds at the intermediate times $S_{i,j}$, $j = 2, \ldots, 6$, as well. Note that the factor $(1-c)^{-1}$ below gives the rate of increase for the number of customers at successive queues. The factor $1 + c^2$ is an error term which we will show to be negligible.

PROPOSITION 4. *Suppose* (5) *holds and that for a given* i, $i \in \{2, \ldots, m\}$,

$$P\left(\xi_{S_{i,1}}(i,1)^- > c^4 M/(1-c)^{i-2}\right) \leq \exp(-\varepsilon_{i,1} M), \tag{32}$$

$$P\left(\left|\xi_{S_{i,1}}(i,1) - M/(1-c)^{i-2}\right| > c\left(\frac{1+c^2}{1-c}\right)^{i-2} M\right) \leq \exp(-\varepsilon_{i,2} M) \tag{33}$$

and

$$P\left(\xi_{S_{i,1}}(i,1)^+ + \gamma_{S_{i,1}}(1,3) > 11c^4 M/(1-c)^{i-2}\right) \leq \exp(-\varepsilon_{i,3} M), \tag{34}$$

with $\varepsilon_{i,1}, \varepsilon_{i,2}, \varepsilon_{i,3} > 0$. *Then for appropriate* $\varepsilon_{i,4}, \varepsilon_{i,5} > 0$ *and large enough* M,

$$P\left(\xi_{S_{i,7}}(i,2)^- > 8c^6 M/(1-c)^{i-2}\right) \leq \exp(-\varepsilon_{i,4} M), \tag{35}$$

$$P\left(\left|\xi_{S_{i,7}}(i) - M/(1-c)^{i-1}\right| > c\left(1 + \tfrac{1}{2}c^2\right)\frac{(1+c^2)^{i-2}}{(1-c)^{i-1}} M\right) \leq \exp(-\varepsilon_{i,5} M) \tag{36}$$

and

$$P\left(\xi_{S_{i,7}}(i,7)^+ + \gamma_{S_{i,7}}(1,3) > 11c^4 M/(1-c)^{i-2}\right) \leq \exp(-\varepsilon_{i,3} M) \tag{37}$$

all hold.

Proposition 5 gives similar bounds at times $S_{i,1}$, $i = 3, \ldots, m$.

PROPOSITION 5. *Suppose that* (35)–(37) *hold for a given* i', $i' \in \{2, \ldots, m-1\}$. *Then* (32)–(34) *all hold with* $i = i' + 1$ *for appropriate* $\varepsilon_{i,1}, \varepsilon_{i,2}, \varepsilon_{i,3} > 0$, *and large enough* M.

By (28)–(30) of Proposition 3, (32)–(34) are satisfied for $i = 2$. One can therefore alternately apply Propositions 4 and 5 to obtain the following

bounds for the network at $t = S_{m,7}$. Since $m = [2c^{-1}\log(c^{-1})]$ and $c \leq 1/100$, we may use the estimates

$$|(1-c)^{-m} - c^{-2}| \leq \tfrac{3}{2}c^{-1}\log(c^{-1}), \qquad (1+c^2)^m \leq 1.1. \tag{38}$$

(m was chosen in (4) so that $(1-c)^{-m} \sim c^{-2}$.)

COROLLARY. *Suppose that* (5) *holds. Then for large enough* M,

$$P\left(\xi_{S_{m,7}}(m,2)^- > 9c^4M\right) \leq \exp(-\varepsilon_{m,4}M), \tag{39}$$

$$P\left(|\xi_{S_{m,7}}(m) - c^{-2}M| > \left(2c^{-1}\log(c^{-1})\right)M\right) \leq \exp(-\varepsilon_{m,5}M) \tag{40}$$

and

$$P\left(\xi_{S_{m,7}}(1,3) + \gamma_{S_{m,7}}(1,3) > 12c^2M\right) \leq \exp(-\varepsilon_{m,3}M). \tag{41}$$

Demonstration of Proposition 4. Before proceeding with the proof of Proposition 4, we obtain a better picture of the qualitative nature of Ξ_t. This is supplied by Lemma 4, which will show that at times $t \geq S_{i,1}$, there are not many customers in $(i,1)^-$. These bounds will enable us to conceptualize new customers as moving quickly to the ith queue. We require an upper bound on the number of customers at $(l,1,3)$ at moderate times, in order to avoid any feedback in the system which could delay the service of new customers at $(1,1)$. A first step is Lemma 2, which ensures that the system is quickly able to complete the service of the customers originally (at $t = 0$) in $(2,7)^+$. Denote by R the time at which the last such customer leaves the system. We observe that by (29) and (9),

$$P\left(S_{2,2} - S_{2,1} < \tfrac{1}{2}cM\right) \leq \exp(-\varepsilon_{18}M), \tag{42}$$

for appropriate $\varepsilon_{18} > 0$ and large M.

LEMMA 2. *Assume that* (5) *holds. Then for appropriate* $\varepsilon_{19}, \varepsilon_{20} > 0$ *and large enough* M,

$$P\left(R > S_{2,1} + \tfrac{1}{4}cM\right) \leq \exp(-\varepsilon_{19}M) \tag{43}$$

and

$$P\left(R > S_{2,2} - \tfrac{1}{4}cM\right) \leq \exp(-\varepsilon_{20}M). \tag{44}$$

PROOF. Only those customers already in $(2,6)^+$ at time $S_{2,1}$ can enter $(2,7)^+$ by time $S_{2,2}$. By (30),

$$P\left(\xi_{S_{2,1}}(2,6)^+ > 11c^4M\right) \leq \exp(-\varepsilon_{11}M). \tag{45}$$

Each of these customers needs to be served at most m times at stages with rates c^{-1} ($(i,1)$, $i = 2,\ldots,m$, and $(1,3)$) and $6m$ times at stages with rates δ^{-1} ((i,j), $i = 2,\ldots,m$ and $j = 2,\ldots,7$). By (45) and (9), the amount of time

spent serving these customers until they leave the system is, off of an exponentially small set, at most

$$12m(c + 6\delta)c^4M \leq \left(25c^3 \log(c^{-1})\right)cM \leq \tfrac{1}{8}cM. \tag{46}$$

Also note that by (28), there are typically at most $6c^5M$ customers in $(2,1)^-$ at $S_{2,1}$. By (9), the number of further customers entering the system over $(S_{2,1}, S_{2,1} + \frac{1}{2}cM]$ is typically at most cM. (The exact choice of the right endpoint is not important.) Each customer in these two groups needs to be served at most twice at stages with rates δ^{-1} before entering the second queue. By (9), the amount of time spent serving these customers there is typically at most

$$3\delta(6c^5 + c)M \leq \tfrac{1}{8}cM. \tag{47}$$

Let ρ denote the amount of time spent serving all of the customers at the stages mentioned above. By (46) and (47),

$$P\left(\rho > \tfrac{1}{4}cM\right) \leq \exp(-\varepsilon_{21}M), \tag{48}$$

for $\varepsilon_{21} > 0$ and large M.

Observe that only the above customers can be served by any queue, other than the second, over $(S_{2,1}, S_{2,2} \wedge (S_{2,1} + \frac{1}{2}cM)]$. By (42), this interval typically contains $(S_{2,1}, S_{2,1} + \frac{1}{2}cM]$. It therefore follows from (48) that all of the queues, with the exception of the second, are simultaneously empty at some $t \leq S_{2,1} + \frac{1}{4}cM$ off of a set of exponentially small probability. All customers originally in $(2,7)^+$ must have left the system by this time. This implies (43). The bound in (44) follows upon another application of (42). □

It was mentioned in Section 1 that the analog of Theorem 1 holds even if one shortens (2) so as to allow only four visits to each queue. We remark here that one reason for instead specifying seven visits is the bounds given in (46). For only four visits, one loses the factor c^3, and therefore needs to work harder elsewhere.

Let $\zeta(s,t)$ denote the total number of (lower) customers ever at $(1,3)$ during $[s,t)$; $\zeta(0,t) = \zeta(t)$ in the notation given before Proposition 1. Also, let L denote the time at which the first of the lower customers originally at $(1,1)$ arrives at $(1,3)$. Since by (5), $\xi_0(1,2) + \xi_0(2) \leq c^5M$, one automatically has

$$\zeta(R,L) \leq c^5M. \tag{49}$$

Comparison of (49) with $\xi_0(1,1)^+ \leq c^4M$ in (5) shows that the number of possible customers visiting $(1,3)$ has been reduced by a factor of c, at least over $[R,L)$. In (44), we saw that, typically, $R < S_{2,2}$. We will also show that, typically, $L > S_{m,7}$. So the analog of (49) typically holds over the interval $(S_{2,2}, S_{m,7}]$. Although the substitution of c^5 for c^4 may at first glance appear insubstantial, it is needed for exhibiting sharp upper bounds for $\xi_t(i,1)^-$ [in particular, the term $2c^6M$ in (50)]. Without these bounds, one would not be able to control $\xi_T(1,1)^+$ sufficiently in (6) to prevent the main body of customers (or "mass") from eventually dissipating throughout the system.

This would break down the "singular" (or "clumped") configurations which are needed in order to show the system is transient.

We now apply Lemma 2 and (49) in conjunction with Proposition 1 and the corollary to Proposition 2 to obtain upper bounds for $\xi_t(i,1)^-$, $i \in \{2,\dots,m\}$, over long stretches of time. Here, we set $L_M = L \wedge M^2$ and

$$F = F_2\big(2c^6M; S_{2,2}, L_M\big) \cap \bigcap_{i=3}^{m} F_i\big(2c^6M; S_{i,1}, L_M\big). \tag{50}$$

LEMMA 3. *Assume that* (5) *holds. For appropriate* $\varepsilon_{22} > 0$ *and large enough* M,

$$P(F^c) \le \exp(-\varepsilon_{22}M). \tag{51}$$

PROOF. Since $S_{2,2} \le S_{3,1} \le \cdots \le S_{m,1}$, one has

$$\begin{aligned} F^c \subset{}& F_2^c\big(2c^6M; S_{2,2}, L_M\big) \\ &\cup \left[\bigcup_{i=3}^{m} \big(F_{i-1}\big(2c^6M; S_{i,1}, L_M\big) \cap F_i^c\big(2c^6M; S_{i,1}, L_m\big)\big)\right]. \end{aligned} \tag{52}$$

On account of (44),

$$\begin{aligned} P\big(F_2^c\big(2c^6M; S_{2,2}, L_M\big)\big) \le{}& P\big(F_2^c\big(2c^6M; R + \tfrac{1}{4}cM, L_M\big)\big) \\ &+ \exp(-\varepsilon_{20}M). \end{aligned} \tag{53}$$

One can apply the strong Markov property to the probability on the right side of (53). Letting $\hat{F}_2(\cdot;\cdot,\cdot)$ denote the set corresponding to the process $\hat{\Xi}_t$ for the queueing network with $\hat{\Xi}_t = \Xi_{t+R}$, one can replace the right side by

$$P\big(\hat{F}_2^c\big(2c^6M; \tfrac{1}{4}cM, L_M - R\big)\big) + \exp(-\varepsilon_{20}M).$$

By (43), typically $R \le S_{2,1} + \frac{1}{4}cM$. So by (28) and (9),

$$P\big(\hat{\xi}_0(2,1)^- > cM\big) \le \exp(-\varepsilon_{23}M),$$

for $\varepsilon_{23} > 0$ [where $\hat{\xi}_0(2,1)^- = \xi_R(2,1)^-$]. From (49), $\hat{\zeta}(L_M - R) = \zeta(R, L_M) \le c^5M$. Setting $N_1 = cM$ and $N_2 = c^5M$, it therefore follows from Proposition 1 and (10) that

$$\begin{aligned} &P\big(F_2^c\big(2c^6M; S_{2,2}, L_M\big)\big) \\ &\quad\le M^2\exp\big(-\varepsilon_2 c^5M\big) \\ &\qquad + \exp(-\varepsilon_{20}M) + \exp(-\varepsilon_{23}M) \\ &\quad\le \exp(-\varepsilon_{24}M), \end{aligned} \tag{54}$$

for $\varepsilon_{24} > 0$ and large M.

One can argue in the same basic manner to control the other terms in (52). Letting ${}_i\hat{F}_{i'}(\cdot;\cdot,\cdot)$ denote the set corresponding to the process ${}_i\hat{\Xi}_t$ for the queueing network with ${}_i\hat{\Xi}_t = \Xi_{t+S_{i,1}}$, one has

$$\begin{aligned} &P\big(F_{i-1}\big(2c^6M; S_{i,1}, L_M\big) \cap F_i^c\big(2c^6M; S_{i,1}, L_M\big)\big) \\ &\quad= P\big({}_i\hat{F}_{i-1}\big(2c^6M; L_M - S_{i,1}\big) \cap {}_i\hat{F}_i^c\big(2c^6M; L_M - S_{i,1}\big)\big). \end{aligned} \tag{55}$$

Note that ${}_i\hat{\xi}_0(i-1) = \xi_{S_{i,1}}(i-1) = 0$ for $i \geq 3$. Setting $N = 2c^6M$, it follows from the corollary to Proposition 2 that

$$P\big(F_{i-1}(2c^6M; S_{i,1}, L_M) \cap F_i^c(2c^6M; S_{i,1}, L_M)\big) \leq M^2 \exp\big(-2\varepsilon_8 c^6 M\big), \tag{56}$$

for large M. Applying (54) and (56) to (52), one obtains

$$P(F^c) \leq \exp(-\varepsilon_{24} M) + mM^2 \exp\big(-2\varepsilon_8 c^6 M\big) \leq \exp(-\varepsilon_{22} M),$$

for appropriate $\varepsilon_{22} > 0$ and large M. □

In order for the bound in (51) to be applicable, the upper bound L_M in (50) must be made more concrete. To do so, we employ the induction assumptions (32)–(34), which we assume hold for a given i, $i \in \{2, \ldots, m\}$. It is important to note here the following feature of the network: All of the M_u customers originally at $(u, 1, 1)$ have by time $S_{1,1}$ advanced beyond $(1, 2)$. On the other hand, all of the M_l customers originally at $(l, 1, 1)$ are still at $(1, 2)$ at $t = S_{1,1}$. All of the lower customers originally at $(1, 1)$ have thus fallen behind all of the corresponding upper customers, and are served at each (i, j), $i = 2, \ldots, m$, only after all such upper customers have been served there. [This is the purpose of the extra stage $(l, 1, 2)$ in (2).] This relationship continues until the upper customers leave the network. In particular, none of the lower customers originally at $(1, 1)$ can arrive at $(1, 3)$ until at least M_u $(\geq \frac{1}{4}M)$ customers have left the network.

LEMMA 4. *Suppose that* (5) *is satisfied. If* (32)–(34) *hold for* $i = 2$, *then for appropriate* $\varepsilon_{2,6} > 0$ *and large enough* M,

$$P\big(\xi_t(2,1)^- > 2c^6 M \text{ for some } t \in [S_{2,2}, S_{2,7}]\big) \leq \exp(-\varepsilon_{2,6} M). \tag{57}$$

If (32)–(34) *hold for a given* $i \in \{3, \ldots, m\}$, *then for appropriate* $\varepsilon_{i,6} > 0$ *and large enough* M,

$$P\left(\xi_t(i,1)^- > 2c^6\left(\frac{1+c^2}{1-c}\right)^{i-2} M \text{ for some } t \in [S_{i,1}, S_{i,7}]\right) \leq \exp(-\varepsilon_{i,6} M). \tag{58}$$

PROOF. By (50) and (51),

$$P\big(F_2^c(2c^6M; S_{2,2}, L_M)\big) \leq \exp(-\varepsilon_{22} M) \tag{59}$$

for large M. To demonstrate (57), it suffices to replace the term L_M $(= L \wedge M^2)$ by $S_{2,7}$. It follows immediately from (34) that

$$P\big(\xi_{S_{2,7}}(2,7)^+ + \gamma_{S_{2,7}}(1,3) > 11c^4 M\big) \leq \exp(-\varepsilon_{2,3} M).$$

So, typically at most $11c^4M < \frac{1}{4}M$ customers have left the network by $t = S_{2,7}$, in which case $L > S_{2,7}$. That is,

$$P(S_{2,7} \geq L) \leq \exp(-\varepsilon_{2,3} M). \tag{60}$$

It follows, on the other hand, from (32)–(34), that

$$P\big(\xi_{S_{2,1}} > 2M\big) \le \exp(-\varepsilon_{2,7}M),$$

with $\varepsilon_{2,7} > 0$ and M large. Also, off of an exceptional set, at most $2M$ further customers enter the network over $(S_{2,1}, S_{2,1} + M]$. Each customer needs to be served at most seven more times at the second queue. So, by (9), the amount of time spent serving all of these customers at the second queue is typically at most $30cM \le \frac{1}{2}M$. So the second queue is empty by then, and $S_{2,7} \le S_{2,1} + \frac{1}{2}M$. Together with the bound for $S_{2,1}$ given in (21), this implies that

$$P(S_{2,7} \ge M) \le \exp(-\varepsilon_{2,8}M), \tag{61}$$

for $\varepsilon_{2,8} > 0$. Plugging (60) and (61) into (59), one obtains

$$P\big(F_2^c(2c^6M; S_{2,2}, S_{2,7})\big) \le \exp(-\varepsilon_{2,6}M)$$

for $\varepsilon_{2,6} > 0$ and large M, as in (57).

The demonstration of (58) is the same, except that one is not presented with a bound for $S_{i,1}$. One can remedy this by applying (32)–(34) and (38) to check that

$$P\big(\xi_{S_{i,1}} + \gamma_{S_{i,1}}(1,3) \ge 2c^{-2}M\big) \le \exp(-\varepsilon_{i,7}M),$$

for $\varepsilon_{i,7} > 0$, and so by (9),

$$P\big(S_{i,1} \ge 3c^{-2}M\big) \le \exp(-\varepsilon_{i,9}M),$$

for $\varepsilon_{i,9} > 0$. One can then check as before that

$$P\big(S_{i,7} \ge 4c^{-2}M\big) \le \exp(-\varepsilon_{i,8}M),$$

for $\varepsilon_{i,8} > 0$, which is the analog of (61). □

The bounds in (57) and (58) are improvements over (32). Together with (33) and (34), they can be employed to demonstrate Proposition 4. The proof consists of obtaining explicit bounds for Ξ_t at the successive times $t = S_{i,j}$, $j = 2, \ldots, 7$, as in (62) and (63). For this, one employs the evolution of Ξ_t over the cycles $(S_{i,j}, S_{i,j+1}]$, $j = 1, \ldots, 6$, together with (9). We note that here, as in Lemma 2, the assumption in (2) of seven rather than fewer visits to each queue simplifies the reasoning.

PROOF OF PROPOSITION 4. It follows immediately from (34) that

$$P\Big(\xi_{S_{i,j}}(i,j)^+ + \gamma_{S_{i,j}}(1,3) > 11c^4M/(1-c)^{i-2}\Big) \le \exp(-\varepsilon_{i,3}M) \tag{62}$$

for $j = 2, \ldots, 7$. Setting $j = 7$ produces (37). To demonstrate (35) and (36), we will employ the bounds

$$P\Big(|\xi_{S_{i,j}}(i,j') - c^{j-j'}M/(1-c)^{i-2}| > n_1(i,j,j')M\Big) \le \exp(-\varepsilon_{i,j-j',1}M) \tag{63}$$

with

$$n_1(i,j,j') = \left(c^{j-j'+1} + \left((5(j-j')c^{(j-j'+3)\wedge 7}) \vee 3c^6\right)\right)\left(\frac{1+c^2}{1-c}\right)^{i-2},$$

for $j = 1,\ldots,7$ and $j' \le j$, and some $\varepsilon_{i,j-j',1} > 0$. [The unwieldy format for $n_1(i,j,j')$ is due to the somewhat different behavior of $\xi_{S_{i,j}}(i,j')$ for small and large $j-j'$; $c^{j-j'+1}$ is the larger term for $j-j' \le 4$.] Application of (57)–(58), together with (63) for $j = 7$ and $j' = 1$, implies (35). (Use (38) to get rid of the factor $(1+c^2)^{i-2}$.) Adding up the estimates in (63) for $j = 7$ and $j' \le 7$, it is also not hard to check that (36) will hold, since the term $\frac{1}{2}c^2$ is easily large enough to absorb the term in $n_1(i,7,j')$ which follows $c^{8-j'}$.

For $j = j' = 1$, (63) follows from (33). We will show that if (63) holds for $j \le 6$ and $j' \le j$, then it holds for $j+1$ and $j' \le j+1$. First note that (63), for the pair $\langle j,j'\rangle$, immediately implies (63) for $\langle j+1, j'+1\rangle$, since the customers under consideration advance precisely one stage over $(S_{i,j}, S_{i,j+1}]$. In order to demonstrate (63) for $\langle j+1,1\rangle$, we first estimate $S_{i,j+1} - S_{i,j}$. Applying (63) at $\langle j,j'\rangle$, $j' \le j$, one can check that the total number of customers in all stages of the ith queue other than the first is typically at most $2M/(1-c)^{i-2}$. Each of these customers is served at rate $\delta^{-1} \ge c^{-8}$, so by (9) the time spent serving them is typically at most $c^7M/(1-c)^{i-2}$. The customers at $(i,1)$ are served at rate c^{-1}. Applying (9) to the bound for $\langle j,1\rangle$ in (63), one can check that the time spent serving all the customers in the ith queue therefore satisfies

$$P\left(|S_{i,j+1} - S_{i,j} - c^jM/(1-c)^{i-2}| > n_2(i,j)M\right) \le \exp(-\varepsilon_{i,j,2}M) \tag{64}$$

with

$$n_2(i,j) = \left(c^{j+1} + (5j-3)c^{(j+3)\wedge 7}\right)\left(\frac{1+c^2}{1-c}\right)^{i-2},$$

for appropriate $\varepsilon_{i,j,2} > 0$. Another application of (9) shows that the number of customers entering the network over $(S_{i,j}, S_{i,j+1}]$ satisfies

$$\begin{aligned} &P\left(|\beta_{S_{i,j+1}}(1,1) - \beta_{S_{i,j}}(1,1) - c^jM/(1-c)^{i-2}| > n_3(i,j)M\right) \\ &\quad \le \exp(-\varepsilon_{i,j,3}M) \end{aligned} \tag{65}$$

with

$$n_3(i,j) = \left(c^{j+1} + (5j-2)c^{(j+3)\wedge 7}\right)\left(\frac{1+c^2}{1-c}\right)^{i-2},$$

for appropriate $\varepsilon_{i,j,3} > 0$.

To justify (63) for $\langle j+1,1\rangle$, we note that all the customers entering the network over $(S_{i,j}, S_{i,j+1}]$ must be in either $(i,1)^-$ or $(i,1)$ at $t = S_{i,j+1}$. Moreover, only those customers entering over $(S_{i,j}, S_{i,j+1}]$ or those in $(i,1)^-$ at $t = S_{i,j}$ can be at $(i,1)$ at $t = S_{i,j+1}$. We can apply (65) together with

(57)–(58) in all cases except for the upper bound when $i = 2$ and $j = 1$, in which case we use (65) and (28). One can check that of the exceptional sets,

$$|\xi_{S_{i,j+1}}(i,1) - c^j M/(1-c)^{i-2}| \le (c^{j+1} + 5jc^{j+3})\left(\frac{1+c^2}{1-c}\right)^{i-2} M \quad \text{for } j \le 3,$$

$$\le (c^{j+1} + 3c^6)\left(\frac{1+c^2}{1-c}\right)^{i-2} M \qquad \text{for } j > 3.$$

This implies (63) for $\langle j+1, 1\rangle$. Hence (63) for $\langle j+1, j'\rangle$, $j' \le j+1$, holds. Induction on j shows that (63) holds for $j \le 7$ and $j' \le j$. This completes the proof of the proposition. □

Demonstration of Proposition 5. The argument for Proposition 5 is considerably shorter than that for Proposition 4. The main estimate is the following lemma, which gives an upper bound for $S_{i+1,1} - S_{i,7}$. The corresponding bound for $S_{1,2} - S_{m,7}$ will be used in Section 5. We therefore, with some abuse of notation, set $S_{m+1,1} = S_{1,2}$ and $(m+1, 1) = (1, 3)$ here, and demonstrate both bounds together.

LEMMA 5. *Assume that* (35)–(37) *hold for a given* $i \in \{2, \dots, m\}$. *Then for appropriate* $\varepsilon_{i,10} > 0$ *and large enough* M,

$$P\left(S_{i+1,1} - S_{i,7} > 10c^7 M/(1-c)^{i-1}\right) \le \exp(-\varepsilon_{i,10} M). \tag{66}$$

PROOF. We use reasoning analogous to that in Lemmas 2 and 4. The only customers served at the ith queue over $(S_{i,7}, S_{i,7} + 11c^7 M/(1-c)^{i-1}]$ are those in $(i+1, 1)^-$ at $t = S_{i,7}$ and those entering the network over this time interval. Each such customer needs to be served at most six times at stages with rates δ^{-1}, and, if starting in $(i, 2)^-$, once at $(i, 1)$ at rate c^{-1}. By (35) and (36), there are typically at $t = S_{i,7}$ at most $2M/(1-c)^{i-1}$ customers in $(i+1, 1)^-$, and by (35), at most $8c^6 M/(1-c)^{i-2}$ customers in $(i, 2)^-$. By (9), there are typically at most $12c^7 M/(1-c)^{i-1}$ customers entering the network. So, typically, at most $3M/(1-c)^{i-1}$ customers need to be served at the ith queue, and $9c^6 M/(1-c)^{i-1}$ customers at $(i, 1)$. Applying (9) again, one obtains that the total time these customers will be served at the ith queue is, off the exceptional set, at most $10c^7 M/(1-c)^{i-1}$, since $c \le 1/100$. The ith queue must in such cases be empty within time $10c^7 M/(1-c)^{i-1}$ of $S_{i,7}$. This implies (66). □

Customers enter the network at rate 1; they are served at $(i+1, 1)$, $i = 2, \dots, m-1$, at rate c^{-1}. Two immediate consequences of Lemma 5 and (9) are therefore given by the following corollary.

COROLLARY. *Assume that* (35)–(37) *hold for a given* $i \in \{2, \dots, m-1\}$. *Then for appropriate* $\varepsilon_{i,11}, \varepsilon_{i,12} > 0$ *and large enough* M,

$$P\left(\beta_{S_{i+1,1}}(1,1) - \beta_{S_{i,7}}(1,1) > 11c^7 M/(1-c)^{i-1}\right) \le \exp(\varepsilon_{i,11} M) \tag{67}$$

and

$$P\Big(\gamma_{S_{i+1,1}}(i+1,1) - \gamma_{S_{i,7}}(i+1,1) > 11c^6M/(1-c)^{i-1}\Big) \leq \exp(-\varepsilon_{i,12}M). \tag{68}$$

Using (67) and (68), the argument for Proposition 5 is now straightforward.

PROOF OF PROPOSITION 5. Only those customers in $(i',1)^-$ at $t = S_{i',7}$, or those entering the system over $(S_{i',7}, S_{i'+1,1}]$ can be in $(i'+1,1)^-$ at $t = S_{i'+1,1}$, since the i'th queue is empty. Together, (35) and (67) provide an easy upper bound for (32), with $i = i'+1$. One has

$$\begin{aligned}\xi_{S_{i'+1,1}}(i'+1,1)^+ + \gamma_{S_{i'+1,1}}(1,3) &\leq \Big(\xi_{S_{i',7}}(i',7)^+ + \gamma_{S_{i',7}}(1,3)\Big) \\ &\quad + \Big(\gamma_{S_{i'+1,1}}(i'+1,1) - \gamma_{S_{i',7}}(i'+1,1)\Big).\end{aligned} \tag{69}$$

That is, the only customers who can be past $(i'+1,1)$ at $t = S_{i'+1,1}$ are those already past $(i',7)$ at $t = S_{i',7}$ and those leaving $(i'+1,1)$ over $(S_{i',7}, S_{i'+1,1}]$. By (37), (68) and (69), (34) holds with $i = i'+1$.

It remains to demonstrate (33) for $i = i'+1$. Note that since the i'th queue is empty at $t = S_{i'+1,1}$,

$$\xi_{S_{i'+1,1}}(i'+1,1) \geq \xi_{S_{i',7}}(i') - \Big(\gamma_{S_{i'+1,1}}(i'+1,1) - \gamma_{S_{i',7}}(i'+1,1)\Big).$$

The lower bound in (33) therefore follows from (36) and (68). Since customers at $(i'+1,1)$ at $t = S_{i'+1,1}$ must come from somewhere in the network at $t = S_{i',7}$ or enter the network over $(S_{i',7}, S_{i'+1,1}]$, an upper bound for the quantity $\xi_{S_{i'+1,1}}(i'+1,1)$ is provided by summing up the bounds in (35)–(37) and (67). Since $c \leq 1/100$, it is easy to see that

$$8c^6(1-c) + c\Big(1 + \tfrac{1}{2}c^2\Big)(1+c^2)^{i'-2} + 11c^4(1-c) + 11c^7 \leq c(1+c^2)^{i'-1}.$$

So the upper bound in (33) with $i = i'+1$ also holds. □

5. Conclusion. The bounds (39)–(41) of the corollary to Propositions 4 and 5 provide us with an accurate description of $\Xi_{S_{m,7}}$. Further detail is given by Lemmas 6 and 7 below. We next analyze $\Xi_{S_{1,2}}$, $S_{1,2}$ being the next time after $S_{m,7}$ at which the mth queue is empty. We then proceed to Ξ_T, T being the time at which the last customer in the first queue at $t = S_{1,2}$ is served. The assertion in (6) of Theorem 2 follows. We then demonstrate (7).

We will use the following bound on $S_{m,7}$.

LEMMA 6. *Assume that* (5) *holds. Then for appropriate* $\varepsilon_{25} > 0$ *and large* M,

$$P\Big(|S_{m,7} - c^{-2}M| > \big(3c^{-1}\log(c^{-1})\big)M\Big) \leq \exp(-\varepsilon_{25}M). \tag{70}$$

PROOF. By (39)–(41), the total number of customers visiting the network by $t = S_{m,7}$ will typically be close to $c^{-2}M$, that is

$$P\Big(|\xi_{S_{m,7}} + \gamma_{S_{m,7}}(1,3) - c^{-2}M| > \big(\tfrac{7}{3}c^{-1}\log(c^{-1})\big)M\Big) \le \exp(-\varepsilon_{26}M),$$

for $\varepsilon_{26} > 0$ and M large. By (5), there were initially between M and $(1 + c^4)M$ customers in the network. So,

$$P\Big(|\beta_{S_{m,7}}(1,1) - c^{-2}M| > \big(\tfrac{8}{3}c^{-1}\log(c^{-1})\big)M\Big) \le \exp(-\varepsilon_{26}M). \tag{71}$$

(70) follows from (71) and (9). □

To demonstrate (6), we will need to bound the number of customers in $(3,1)^-$.

LEMMA 7. *Assume that* (5) *holds. Then for large* M,

$$P(S_{m,7} \ge L) \le \exp(-\varepsilon_{m,3}M) \tag{72}$$

and

$$P\Big(\xi_{S_{m,7}}(3,1)^- > 2c^6M\Big) \le \exp(-\varepsilon_{27}M) \qquad \text{for } \varepsilon_{27} > 0. \tag{73}$$

PROOF. By (41), typically at most $12c^2M < \frac{1}{4}M$ customers have left the network by $t = S_{m,7}$. As shown after Lemma 3, this implies that on this set, none of the lower customers originally at $(1,1)$ has yet arrived at $(1,3)$, and so $S_{m,7} < L$. This implies (72). By (70), $S_{m,7} < M^2$ is typically also the case. So $S_{m,7} < L \wedge M^2 = L_M$ typically holds. By (50) and (51),

$$P\big(F_3^c\big(2c^6M; S_{3,1}, L_M\big)\big) \le \exp(-\varepsilon_{22}M),$$

and (73) follows by substituting $S_{m,7}$ for L_M. □

We mention several simple consequences of Lemma 5, with $i = m$, and of Lemma 6 regarding the behavior of Ξ_t up to $t = S_{1,2}$. Since customers enter the network at rate 1, it follows from Lemma 5, (38) and (9) that

$$P\Big(\beta_{S_{1,2}}(1,1) - \beta_{S_{m,7}}(1,1) > 11c^5M\Big) \le \exp(-\varepsilon_{28}M), \tag{74}$$

for $\varepsilon_{28} > 0$ and M large. Upper customers exit from the network after $(m,7)$, but lower customers are served at $(1,3)$ at rate c^{-1}. So one also obtains from Lemma 5 that

$$P\Big(\gamma_{S_{1,2}}(l,1,3) - \gamma_{S_{m,7}}(l,1,3) > 11c^4M\Big) \le \exp(-\varepsilon_{29}M), \tag{75}$$

for $\varepsilon_{29} > 0$ and M large. By the two lemmas,

$$P\big(|S_{1,2} - c^{-2}M| > \big(4c^{-1}\log(c^{-1})\big)M\big) \le \exp(-\varepsilon_{30}M), \tag{76}$$

for $\varepsilon_{30} > 0$ and large M. Together with (9), (76) allows us to control the number of lower customers entering the network by $t = S_{1,2}$, so that

$$P\Big(|\beta_{S_{1,2}}(l,1,1) - \tfrac{1}{2}c^{-2}M| > \big(\tfrac{5}{2}c^{-1}\log(c^{-1})\big)M\Big) \le \exp(-\varepsilon_{31}M), \tag{77}$$

for $\varepsilon_{31} > 0$ and large M.

We now analyze the behavior of $\Xi_{S_{1,2}}$.

PROPOSITION 6. *Assume that* (5) *holds. Then for appropriate* $\varepsilon_k > 0$ *and large* M,

$$P\left(\xi_{S_{1,2}}(1,3)^- > 10c^4M\right) \le \exp(-\varepsilon_{32}M), \tag{78}$$

$$P\left(\xi_{S_{1,2}}(3,1)^- > 12c^5M\right) \le \exp(-\varepsilon_{33}M), \tag{79}$$

$$P\left(|\xi_{S_{1,2}}(l,1,3) - \tfrac{1}{2}c^{-2}M| > \left(3c^{-1}\log(c^{-1})\right)M\right) \le \exp(-\varepsilon_{34}M) \tag{80}$$

and

$$P\left(\gamma_{S_{1,2}}(l,1,3) > 12c^4M\right) \le \exp(-\varepsilon_{35}M). \tag{81}$$

PROOF. Only those customers in $(m,1)^-$ at $t = S_{m,7}$, or those entering the network over $(S_{m,7}, S_{1,2}]$ can be in $(1,3)^-$ at $t = S_{1,2}$, since the mth queue is empty. So (78) follows from (39) and (74). Similarly, only those customers already in $(3,1)^-$ or those entering the network over $(S_{m,7}, S_{1,2}]$ can be in $(3,1)^-$ at $t = S_{1,2}$, and so (79) follows from (73) and (74). By (5), $\xi_0(1,1)^+ \le c^4M$. Together with (72), this implies that

$$P\left(\gamma_{S_{m,7}}(l,1,3) > c^4M\right) \le \exp(-\varepsilon_{m,3}M). \tag{82}$$

Inequality (81) follows from (75) and (82).

To demonstrate (80), we note that by (77), typically about $\frac{1}{2}c^{-2}M$ lower customers have entered the network by $t = S_{1,2}$. Since these customers must be somewhere, subtraction of the bounds given in (78) and (81) from (77) gives the lower bound for (80). Since at $t = 0$, there are less than M lower customers in the network, the upper bound in (80) also follows from (77). □

$T - S_{1,2}$ is the time required for all the customers in the first queue at $t = S_{1,2}$ to be served. Customers at $(1,1)$ and $(1,2)$ are served at rate δ^{-1}, and those at $(1,3)$, at rate c^{-1}. Applying (9) to (78) and (80), it is easy to check that

$$P\left(|T - S_{1,2} - \tfrac{1}{2}c^{-1}M| > \left(\tfrac{7}{2}\log(c^{-1})\right)M\right) \le \exp(-\varepsilon_{36}M), \tag{83}$$

for $\varepsilon_{36} > 0$ and large M. Another application of (9) implies that

$$\begin{aligned} &P\left(|\beta_T(u,1,1) - \beta_{S_{1,2}}(u,1,1) - \tfrac{1}{4}c^{-1}M| > \left(2\log(c^{-1})\right)M\right) \\ &\qquad \le \exp(-\varepsilon_{37}M) \end{aligned} \tag{84}$$

and

$$\begin{aligned} &P\left(|\beta_T(l,1,1) - \beta_{S_{1,2}}(l,1,1) - \tfrac{1}{4}c^{-1}M| > \left(2\log(c^{-1})\right)M\right) \\ &\qquad \le \exp(-\varepsilon_{37}M), \end{aligned} \tag{85}$$

for $\varepsilon_{37} > 0$ and large M. It is now straightforward to analyze the behavior of Ξ_T.

PROPOSITION 7. *Assume that* (5) *holds. Then for large enough* M,

$$P\big(|\xi_T(u,1,1) - \tfrac{1}{4}c^{-1}M| > \big(2\log(c^{-1})\big)M\big) \le \exp(-\varepsilon_{37}M), \tag{86}$$

$$P\big(|\xi_T(l,1,1) - \tfrac{1}{4}c^{-1}M| > \big(2\log(c^{-1})\big)M\big) \le \exp(-\varepsilon_{37}M), \tag{87}$$

$$P\big(\xi_T(1,1)^+ > 10c^4M\big) \le \exp(-\varepsilon_{32}M) \tag{88}$$

and

$$P\big(\xi_T(1,2) + \xi_T(2) > 12c^5M\big) \le \exp(-\varepsilon_{33}M), \tag{89}$$

where ε_{32}, ε_{33} *and* ε_{37} *are as in* (78), (79) *and* (84)–(85).

PROOF. Over $(S_{2,1}, T]$, those customers arriving in the network end up at $(1,1)$, those starting at $(1,3)$ leave the network, and those starting elsewhere in the network either leave the network or end up in $(1,1)^+$. Consequently, (86) follows immediately from (84), (87) from (85), (88) from (78), and (89) from (79). □

The bounds in (6) of Theorem 2 follow easily from Proposition 7, since $c \le 1/100$. By decreasing c further, they can of course be improved. By replacing (84)–(85) with bounds for $\beta_T(u,1,1) - \beta_{S_{1,2}}(u,1,1)$ relative to $\beta_T(1,1) - \beta_{S_{1,2}}(1,1)$, one can choose $\xi_T(u,1,1)/\xi_T(1,1)$ as close to $\frac{1}{2}$ as desired. Also note that by (76) and (83),

$$P\big(|T - c^{-2}M| > \big(5c^{-1}\log(c^{-1})\big)M\big) \le \exp(-\varepsilon_{38}M), \tag{90}$$

for appropriate $\varepsilon_{38} > 0$ and large M.

We still need to demonstrate (7) of Theorem 2.

PROPOSITION 8. *Assume that* (5) *holds. Then for appropriate* $\varepsilon_{39} > 0$ *and large enough* M,

$$P\big(\xi_t < \tfrac{1}{5}M \text{ for some } t \in [0,T]\big) \le \exp(-\varepsilon_{39}M). \tag{91}$$

PROOF. By (81), the number of lower customers leaving the network by $t = S_{1,2}$ is typically at most $12c^4M$. Since $M_l \ge \frac{1}{4}M$, it follows that

$$P\big(\xi_t < \tfrac{1}{5}M \text{ for some } t \in [0, S_{1,2}]\big) \le \exp(-\varepsilon_{35}M). \tag{92}$$

(The bound can be made arbitrarily close to M by decomposing $[0, S_{1,2}]$.) To examine the behavior of ξ_t over $(S_{1,2}, T]$, we introduce the time $T' = S_{1,2} + 2M$. By (9),

$$P\big(\beta_{T'}(1,1) - \beta_{S_{1,2}}(1,1) < M\big) \le \exp(-\varepsilon_{40}M) \tag{93}$$

and

$$P\big(\gamma_{T'}(l,1,3) - \gamma_{S_{1,2}}(l,1,3) > 4c^{-1}M\big) \le \exp(-\varepsilon_{41}M), \tag{94}$$

for $\varepsilon_{40}, \varepsilon_{41} > 0$ and large M. On account of (80), there are typically about $\frac{1}{2}c^{-2}M$ customers at $(l, 1, 3)$ at $t = S_{1,2}$. Together with (94), this shows that

$$P\big(\xi_t < M \text{ for some } t \in (S_{1,2}T']\big) \leq \exp(-\varepsilon_{42}M), \tag{95}$$

for $\varepsilon_{42} > 0$ and large M. Since customers entering the network after $S_{1,2}$ remain at $(1, 1)$ until time T, (93) implies that

$$P\big(\xi_t < M \text{ for some } t \in (T' \wedge T, T]\big) \leq \exp(-\varepsilon_{40}M). \tag{96}$$

Inequality (91) follows from (92), (95) and (96). □

Proposition 8 demonstrates (7), which completes the proof of Theorem 2. Consequently, Theorem 1 also holds. The process Ξ_t is thus unstable, with $\xi_t \to \infty$ as $t \to \infty$. More detail on the asymptotics of Ξ_t under (5) is provided by Propositions 4–7.

REFERENCES

[1] BRAMSON, M. (1994). Instability of FIFO queueing networks. *Ann. Appl. Probab.* **4** 414–431.
[2] GLYNN, P. W. AND MEYN, S. P. (1994). A Lyapunov bound for solutions of Poisson's equation. *Ann. Probab.* To appear.
[3] KELLY, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley, New York.
[4] LU, S. H. and KUMAR, P. R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Automat. Control* **36** 1406–1416.
[5] RYBKO, S. and STOLYAR, A. L. (1992). Ergodicity of stochastic processes that describe functioning of open queueing networks. *Problems Inform. Transmission* **28** 3–26 (in Russian).
[6] SEIDMAN, T. I. (1994). "First come, first served" can be unstable! *IEEE Trans. Automat. Control*. To appear.
[7] SEIDMAN, T. I. and YERSHOV, A. (1994). A single product instability example for FCFS. Unpublished manuscript.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF WISCONSIN–MADISON
VAN VLECK HALL
480 LINCOLN DRIVE
MADISON, WISCONSIN 53706