The Annals of Applied Probability
1999, Vol. 9, No. 3, 818–853

A STABLE QUEUEING NETWORK WITH UNSTABLE FLUID MODEL[1]

BY MAURY BRAMSON

University of Minnesota

Fluid models have become a standard tool for demonstrating stability for queueing networks. It is presently not known, however, when the stability of a fluid model follows from that of the corresponding queueing network. We present an example of a queueing network where such stability does not, in fact, follow. This example also shows that the behavior of the fluid limits and the fluid model solutions for the same queueing network can differ considerably from one another.

1. Introduction. There has recently been considerable interest in the qualitative behavior of open multiclass queueing networks. An important question in this context is whether a given queueing network is stable, that is, its underlying Markov process is positive recurrent. Various examples have shown that there is no general criterion for this behavior [see, e.g., Bramson (1994), Lu and Kumar (1991), Rybko and Stolyar (1992) and Seidman (1994)]. One therefore needs to examine such networks on a case-by-case basis.

The standard approach for investigating the stability of a queueing network is the analysis of the associated fluid limits. These are the different "limits" one obtains by shrinking the weight of individual customers and time proportionally. The fluid limits will satisfy fluid model equations, which correspond to the deterministic analog of the queueing network under consideration. Typically, one attempts to show that solutions of the fluid model equations are stable, that is, their queue lengths are 0 by a fixed time. The stability of the queueing network then follows from the stability of these solutions [Dai (1995)].

The importance of fluid limits and fluid model equations in showing the stability of queueing networks has led to questions about the converse direction. Namely, does the stability of the fluid limits or of the fluid model follow from that of the associated queueing network? Little is known in this direction, with the only results stating, in effect, that when the fluid limits all have a uniformly positive drift, then the queueing network itself is unstable [Dai (1996), Meyn (1995)].

We present here a family of queueing networks that are stable, but whose fluid models are unstable, that is, there exists an unstable solution of the fluid model equations. What occurs, in essence, is that the random oscillations

Received February 1998; revised November 1998.

[1] Supported in part by NSF Grant DMS-96-26196.

AMS 1991 *subject classifications*. 60K25, 68M20, 90B22.

Key words and phrases. Queueing networks, fluid limits, fluid models.

appearing in the queueing network prevent its fluid limits from becoming trapped among certain unstable solutions of the associated fluid model that grow linearly in time. Our example consists of networks which can be thought of as "almost" being reentrant lines consisting of three stations, each with two classes, and an assigned priority. The actual networks are formed by replacing the middle station by a family of identical two-class stations that lie in parallel with one another. All of the service and interarrival times of the networks are exponentially distributed.

It is routine to show that the fluid models associated with the above queueing networks are unstable. It is, however, more difficult to show that the networks themselves are stable. One needs to do detailed bookkeeping following the evolution of the process, in order to show that the fluid limits are eventually 0 at the above middle stations. The fluid models one obtains by deleting these stations are last-buffer-first-served, and are easily shown to be stable. Applying a variant of the stability result in Dai (1995), this leads to the desired stability for the original queueing networks. The reasoning employed here also shows that the fluid limits associated with the original networks are asymptotically stable. (This property is slightly weaker than stability and is easier to work with in our setting.) Thus, the fluid limits and the fluid models for these networks also differ in behavior.

The paper is organized as follows. In Section 2, we introduce terminology for the queueing networks and fluid models, and present our results. Theorem 1 states that the queueing networks we consider are stable, and Theorem 2 shows that the associated fluid models are not. Theorem 2 is demonstrated in Section 3. The proof of Theorem 1 constitutes the remaining seven sections of the paper. In Section 4, the problem is rephrased, in Theorem 3, in terms of the asymptotic stability of fluid limits. An outline of the following sections is then given. In Section 5, certain large deviation bounds are given. Sections 6–9 are devoted to deriving upper bounds on the number of customers at the middle stations of the queueing network, which were alluded to earlier. In Sections 6 and 7, bounds are given on the growth of the number of customers at the classes having higher priority; in Section 8, bounds are given at the classes having lower priority. Using the strict subcriticality of these stations, it is shown in Section 9 that their classes frequently empty in a "coordinated" manner. Section 10 employs the results from these sections to show that the fluid limits at these stations are eventually 0. Using known results on last-buffer-first-served networks, Theorem 3 is then demonstrated.

2. Terminology, results and basic ideas. The queueing networks that we will investigate consist of $L+2$ stations, with each station possessing two classes (or buffers). We will distinguish classes at a given station by the letters a and b; since L of the stations will perform a similar role, we designate the stations in the network by 1, 2, $(3,1)$, $(3,2)$, $\dots$, $(3,L)$. For a given class at a station, we write, for example, $(3,2,a)$. Customers at the class a at a station will be assumed to always have priority over customers at the class b at the same station, and will preempt service at b upon arrival.

All of the service and interarrival times for the queueing network will be assumed to be exponentially distributed and mutually independent. Customers are assumed to enter the network at rate 1 at class $(1, b)$, and are assumed to move through the network until leaving it after visiting class $(1, a)$, by the following routes:

$$\to (1, b) \to (2, b) \to (3, l, a) \to (3, l, b) \to (2, a) \to (1, a) \to, \tag{2.1}$$

where $l = 1, \dots, L$. Thus, the routing is deterministic, with the exception of the choice following departure from $(2, b)$; in this case, we assume that a customer randomly chooses each of the classes $(3, l, a)$, $l = 1, \dots, L$, with equal probability. The mean service times m_k at the different classes k are given by:

$$\begin{gathered} m_{1,a} = \frac{3}{4}, \ m_{1,b} = \gamma, \quad m_{2,a} = \gamma, \quad m_{2,b} = \frac{3}{4}, \\ m_{3,l,a} = \frac{3}{4}L, \ m_{3,l,b} = \frac{\gamma}{L}, \end{gathered} \tag{2.2}$$

where we assume that $\gamma \in (0, 1/8)$. Thus, upon entering the network, each customer first enters the “quick” class $(1, b)$, followed by the “slow” classes $(2, b)$ and $(3, l, a)$, the “quick” classes $(3, l, b)$ and $(2, a)$, and the “slow class” $(1, a)$, before departing. Throughout the paper, γ will be assumed to be fixed, whereas L will be allowed to vary. As above, we will use the index k to denote class level quantities; j will be used to denote station level quantities.

The *total arrival rate* λ_k for each class at each of the first two stations is 1, and is $1/L$ for the classes at the stations $(3, l)$, $l = 1, \dots, L$. The *traffic intensity* ρ_j at a station j is given by $\rho_j = \sum_{k \in \mathscr{C}(j)} m_k \lambda_k$, where $\mathscr{C}(j)$ denotes the set of classes belonging to the station. Here, it is easy to see that

$$\rho_1 = \rho_2 = \frac{3}{4} + \gamma, \qquad \rho_{3,l} = \frac{3}{4} + \frac{\gamma}{L^2}. \tag{2.3}$$

Consequently, $\rho_j < 1$ holds for all j, that is, the queueing network is *strictly subcritical*. When $\rho_j < 1$ holds at a given j, that station is also said to be strictly subcritical.

We will use $Z(t)$ to denote the underlying right continuous Markov process of the queueing network. Its state space can be chosen to be the subset of $\mathbb{Z}^K$, $K = 2(L + 2)$, with nonnegative coordinates. Its transition rates μ_k are given by the reciprocals of the means m_k in (2.2). Since the sample paths are piecewise constant, $Z(t)$ is strong Markov. We will use $Z_j(t)$ to denote the number of customers at the station j, and $Z_k(t)$ to denote the number of customers at the class k, writing, for example, $Z_{2,a}(t)$. We will set $Z_3(t) = \sum_{l=1}^{L} Z_{3,l}$, and define $Z_{3,a}(t)$ and $Z_{3,b}(t)$ analogously. The total number of customers in the system at time t will be denoted by $|Z(t)|$. We let $T(t) = (T_1(t), \dots, T_K(t))$ denote the continuous process of cumulative service times associated with $Z(t)$, and write $T_j(t) = \sum_{k \in \mathscr{C}(j)} T_k(t)$ for the cumulative service time at a station j.

In this paper, we demonstrate two main results about the behavior of the network in (2.1), (2.2). The first is given by Theorem 1, which says that, for large L, the network is stable.

THEOREM 1. *For sufficiently large* L, *the Markov process* $Z(t)$ *of the queueing network given in* (2.1), (2.2) *is positive recurrent.*

Since all states of $Z(t)$ communicate, it follows from the theorem that the process has a unique equilibrium π. Using the machinery from Dai and Meyn (1995), one can show that π possesses all moments. (More detail will be given later.)

Most of the work in this paper consists of demonstrating Theorem 1. The main step will be to show that the corresponding fluid limits are asymptotically stable. (The collection of fluid limits will also be referred to as the "fluid limit model.") A summary of the reasoning employed for Theorem 1, together with the relevant terminology, is given in Section 4. The details for the individual parts comprise Sections 5–10.

Our interest in the queueing network (2.1), (2.2) stems from the fact that although it is stable, the corresponding fluid model is not. The queueing network thus exhibits "borderline" behavior. Before discussing the fluid model, we make several observations which will hopefully help to motivate the choice of the network.

Understanding the queueing network. We first observe what happens when one "collapses" the stations $(3,l)$, $l = 1,\dots,L$, into a single station 3. The corresponding network is then a reentrant line, with entrance rate 1, where all customers follow the route

$$\to (1,b) \to (2,b) \to (3,a) \to (3,b) \to (2,a) \to (1,a) \to, \tag{2.4}$$

and where the mean service times are the same as in (2.2), except that

$$m_{3,a} = \frac{3}{4}, \qquad m_{3,b} = \frac{\gamma}{L^2}. \tag{2.5}$$

The traffic intensities ρ_j are therefore the same as in (2.3).

Superficially, the behavior for this collapsed network is the same as that for the original network, and it is simpler to think in terms of it instead. The main difference is that it is easier for one of the classes $(3,l,a)$ in the original network to become empty than it is for $(3,a)$ in the collapsed network. When either occurs, the effects we are seeking are the same. When $(3,a)$, for instance, is empty, this allows service at the lower priority class $(3,b)$, from which customers pass to $(2,a)$ (which was likely empty). Service there will then interrupt service at the lower priority class $(2,b)$, and hence delay the creation of new $(3,a)$ customers. Since $(3,b)$ customers will continue to be served until $(3,a)$ is again occupied, all of the $(3,b)$ customers are likely to be served before then. This behavior serves to prevent the build up of $(3,b)$ customers in the network. It is important for the analysis of either network.

Dealing now with the collapsed network in (2.4), (2.5), we note that the neighboring classes $(2, b)$ and $(3, a)$ have the same mean service time 3/4. This implies that the number of customers in $(3, a)$ is dominated by a continuous time symmetric random walk with reflection at 0. (The comparison becomes one-sided when there is no service at $(2, b)$, either because $(2, b)$ is empty or there are customers at $(2, a)$.) So, the number of customers there, when scaled diffusively, can be compared with a reflecting Brownian motion. This implies that the number of customers at $(3, a)$ can only grow sublinearly.

We would also like to obtain upper bounds on the time required for the class $(3, a)$ to become empty, and then employ these, in conjunction with the behavior mentioned two paragraphs above, to obtain sublinear upper bounds on the number of customers at the class $(3, b)$. Although we are not able to obtain these bounds for the collapsed network, the better bounds that one has, on when the classes $(3, l, a)$ in the original network empty, suffice. Consequently, the total number of customers at the stations $(3, l)$, $l = 1, \dots, L$, only grows sublinearly.

Since all of the stations in the original network are strictly subcritical, they will each eventually empty (although not simultaneously). The sublinear growth of the number of customers at the stations $(3, l)$, $l = 1, \dots, L$, will allow us to omit these stations entirely from the network without changing its basic behavior. This reduction produces the network with entrance rate 1, where all customers follow the route

$$\rightarrow (1, b) \rightarrow (2, b) \rightarrow (2, a) \rightarrow (1, a) \rightarrow \tag{2.6}$$

and where the mean service times are

$$m_{1,a} = m_{2,b} = \tfrac{3}{4}, \qquad m_{1,b} = m_{2,a} = \gamma, \tag{2.7}$$

with $\gamma \in (0, 1/8)$. This network is last-buffer-first-served, and is known to be stable [Dai and Weiss (1995)]. This same behavior, we claim, is maintained by the original network.

It is helpful to tinker a bit more with the networks we have considered above to see how stability may also fail. It is the role of station 2 to induce the above random walk-like fluctuations at $(3, a)$ for both the original and the collapsed networks. Removing station 2 from the collapsed network produces the network with entrance rate 1 and route

$$\rightarrow (1, b) \rightarrow (3, a) \rightarrow (3, b) \rightarrow (1, a) \rightarrow . \tag{2.8}$$

The mean service times are, as before,

$$m_{1,a} = m_{3,a} = \frac{3}{4}, \qquad m_{1,b} = \gamma, \qquad m_{3,b} = \frac{\gamma}{L^2}. \tag{2.9}$$

This is a variant of the well-known Lu–Kumar network. The rates of service we have chosen here for the different classes ensures that the queueing network given by (2.8), (2.9) will be unstable [Dai and Weiss (1996)]. So, the original network given by (2.1), (2.2) can be thought of as a Lu–Kumar network with additional stations which stabilize the network due to the introduction of random fluctuations.

The fluid model for (2.1), (2.2). We will contrast the stability of the original network in (2.1), (2.2) with the instability of the corresponding fluid model. First, we recall the basics of fluid models. For this, we employ the notation in Dai (1995).

Fluid models are the continuous, deterministic analogs of queueing networks. Here, the notion of customers is replaced by that of mass, which is to be thought of as circulating around the system. Equations are given which tie together the evolution of relevant quantities, such as the vector of queue lengths $\bar{Z}(t) = (\bar{Z}_1(t), \dots, \bar{Z}_K(t))$, $t \geq 0$, of the different classes, and the vector of cumulative service times $\bar{T}(t) = (\bar{T}_1(t), \dots, \bar{T}_K(t))$, $t \geq 0$. We find it convenient to set $\bar{Z}_j(t) = \sum_{k\in\mathscr{C}(j)} \bar{Z}_k(t)$ and $\bar{T}_j(t) = \sum_{k\in\mathscr{C}(j)} \bar{T}_k(t)$. In accordance with the corresponding queueing network quantities, one stipulates that $\bar{Z}_k(t) \geq 0$, $k = 1, \dots, K$, always holds, and that $\bar{T}(0) = 0$, with $\bar{T}_k(t)$, $k = 1, \dots, K$, and $t - \bar{T}_j(t)$, $j = 1, \dots, J$, being nondecreasing. We employ the notation $\bar{Z}_k^+(t)$ to denote the sum of the queue lengths at all classes at the station to which k belongs, which have priority at least as great as k. That is, in our setting,

$$\bar{Z}_{j,a}^+(t) = \bar{Z}_{j,a}(t),$$
$$\bar{Z}_{j,b}^+(t) = \bar{Z}_{j,a}(t) + \bar{Z}_{j,b}(t).$$

The quantities $\bar{T}_k^+(t)$ are defined analogously.

To formulate the fluid model equations (2.10), (2.11) below, we will also find it convenient to use notation which is standard for networks with general routing. The $K \times K$ matrix P denotes the transition matrix between classes. In our case, at all classes k_1 except for $k_1 = (2, b)$, P_{k_1,k_2} is 1 or 0, depending on whether or not class k_2 immediately succeeds k_1 according to the route given in (2.1). For $k_1 = (2, b)$, $P_{k_1,k_2} = 1/L$ for $k_2 = (3, l, a)$, $l = 1, \dots, L$, and $P_{k_1,k_2} = 0$ for other k_2. Denote by P^t the transpose of P. Let $\alpha = (\alpha_1, \dots, \alpha_c)$ be the vector with $\alpha_k = 1$ for $k = (1, b)$ and $\alpha_k = 0$ elsewhere; α gives the rate of exogeneous arrivals into the system. Also, M denotes the $K \times K$ matrix with diagonal entries equal to the mean service times given in (2.2) and other entries equal to 0.

Using the above notation, the *fluid model equations* corresponding to the network given in (2.1), (2.2) are

$$\bar{Z}(t) = \bar{z} + \alpha t + (P - I)^t M^{-1}\bar{T}(t), \tag{2.10}$$

$$t - \bar{T}_k^+(t) \text{ can only increase when } \bar{Z}_k^+(t) = 0, \qquad k = 1, \dots, K, \tag{2.11}$$

where $\bar{z} = \bar{Z}(0)$. [Here, α, $\bar{T}(t)$ and $\bar{Z}(t)$ are written as column vectors.] Solutions $(\bar{T}(t), \bar{Z}(t))$ of (2.10), (2.11) are *fluid model solutions*. Equation (2.10) gives the effect of arrivals and departures at a class upon the queue length there. Since $\bar{T}(t)$ is continuous, the continuity of $\bar{Z}(t)$ follows from (2.10).

Equation (2.11) incorporates the priority scheme for service at each station, with each nonempty class only receiving service when none of the higher-priority classes at the station is occupied. It can be rewritten as

$$\int_0^\infty \bar{Z}_k^+(t)\, d(t - \bar{T}_k^+(t)) = 0, \qquad k = 1, \dots, K. \tag{2.11$'$}$$

One says that a fluid model is *stable* if, for a given $\delta > 0$ and all solutions of the fluid model equations with $|\bar{z}| = 1$, one has $\bar{Z}(t) = 0$ for $t \geq \delta$. (As before, $|\cdot|$ denotes the sum of the coordinates.) The above fluid model equations scale, so this is equivalent to $\bar{Z}(t) = 0$ always holding for $t \geq \delta|\bar{z}|$. An important result is that if a fluid model is stable, then the associated queueing network is stable [Dai (1995) and Stolyar (1994)]. Theorem 1 and Theorem 2, which is given below, imply that the converse is not true. In fact, there exists a solution of (2.10), (2.11), with $\bar{z} = 0$, such that $|\bar{Z}(t)|$ increases linearly to infinity.

THEOREM 2. *Fix L. There exists a solution* $(\bar{T}(t), \bar{Z}(t))$ *of the fluid model equations* (2.10), (2.11), *corresponding to the queueing network given in* (2.1), (2.2), *with* $\bar{z} = 0$ *and*

$$\liminf_{t\to\infty} |\bar{Z}(t)|/t = 1/3. \tag{2.12}$$

We will demonstrate Theorem 2 in the next section by explicitly constructing a solution $(\bar{T}(t), \bar{Z}(t))$ of (2.10), (2.11) which satisfies (2.12). The argument is straightforward and consists of following the flow of mass over a given time interval, which is then pieced together with scaled versions of this flow over other intervals. We note that the presence of a multitude of 3-stations plays no role in the construction of our example, which is also (after a minor change in notation) a solution of the fluid model equations associated with the collapsed network given by (2.4), (2.5). Also, note that for the collapsed network, $m_{2,b} = m_{3,a}$ and $m_{2,a} < m_{1,a}$. For our example, the presence of station 2, therefore, does not slow down the flow of mass into the succeeding classes $(3, a)$ and $(1, a)$. [This is automatic for $(1, a)$, but not for $(3, a)$.] Removal of station 2 from the collapsed network produces the network in (2.8), (2.9). As mentioned there, this last network is an unstable Lu–Kumar network. Viewed in this light, the behavior given in Theorem 2 is not too surprising.

We conclude this section with several comments about the relationships between the mean service times of the different classes of the network in (2.1), (2.2). To obtain the behavior exhibited by our example, one requires that $m_{3,l,a} = Lm_{2,b}$; the same value is assigned to $m_{1,a}$ for simplicity. One needs the means at the classes $(1, b)$ and $(2, a)$ to be smaller and that at $(3, l, b)$ to be much smaller. The value γ is chosen small enough so that all of the stations are strictly subcritical. Moreover, the means at $(1, a)$ and $(3, l, a)$ need to satisfy $m_{1,a} + m_{3,l,a}/L > 1$, so that the corresponding Lu–Kumar network in (2.8), (2.9) is unstable.

3. Instability of the fluid model.

Here, we demonstrate the instability of the fluid model given in Theorem 2. We will show, in particular, that $\liminf_{t\to\infty} |\bar{Z}(t)|/t = 1/3$ for a given solution of the fluid model equations (2.10), (2.11), whose parameters correspond to those for the queueing network in (2.1), (2.2). Our reasoning proceeds in two steps. We first, as in Proposition 3.1, construct a solution $(\bar{T}(t), \bar{Z}(t))$ of the fluid model equations on $[0, 6]$ with $\bar{z} = |\bar{Z}(0)| = 1$ and $|\bar{Z}(6)| = 3$, where all the mass at $t = 0$ and $t = 6$ is concentrated at class $(1, b)$. We then piece together scaled versions of this solution so that the resulting solution is defined over $[0, \infty)$, and grows linearly starting at $\bar{z} = 0$.

We construct the function $(\bar{T}(t), \bar{Z}(t))$, $t \in [0, 6]$, piecewise over the time intervals with endpoints 0, $\gamma/(1-\gamma)$, 3, $3+(4\gamma/L^2)$, $3+4\gamma$ and 6. We assume that $\bar{T}'(t)$ is constant over each of these intervals, proceeding as follows:

$$
\begin{aligned}
\bar{T}'_{1,b}(t) &= \begin{cases} 1, & \text{for } t \in (0, \gamma/(1-\gamma)),\\ \gamma, & \text{for } t \in (\gamma/(1-\gamma), 3),\\ 0, & \text{for } t \in (3, 6),\end{cases}\\
\bar{T}'_{2,b}(t) = \bar{T}'_{3,l,a}(t) &= \begin{cases} 1, & \text{for } t \in (0, 3),\\ 0, & \text{for } t \in (3, 6),\end{cases}\\
\bar{T}'_{3,l,b}(t) &= \begin{cases} 1, & \text{for } t \in (3, 3+4\gamma/L^2),\\ 0, & \text{otherwise},\end{cases}\\
\bar{T}'_{2,a}(t) &= \begin{cases} 1 & \text{for } t \in (3, 3+4\gamma),\\ 0, & \text{otherwise},\end{cases}\\
\bar{T}'_{1,a}(t) &= \begin{cases} 1, & \text{for } t \in (3, 6),\\ 0, & \text{for } t \in (0, 3).\end{cases}
\end{aligned}
\tag{3.1}
$$

Integration then, of course, gives $\bar{T}(t)$. We assume that $\bar{Z}(t)$ is linear over segments consisting of one or more of these time intervals, with values at the endpoints of the segments given by

$$
\begin{aligned}
&\bar{Z}_{1,b}(0) = 1,\ \bar{Z}_{1,b}(\gamma/(1-\gamma)) = \bar{Z}_{1,b}(3) = 0,\ \bar{Z}_{1,b}(6) = 3,\\
&\bar{Z}_{2,b}(0) = 0,\ \bar{Z}_{2,b}(\gamma/(1-\gamma)) = (3-4\gamma)/3(1-\gamma),\\
&\quad \bar{Z}_{2,b}(3) = \bar{Z}_{2,b}(6) = 0,\\
&\bar{Z}_{3,l,a}(0) = \bar{Z}_{3,l,a}(6) = 0,\\
&\bar{Z}_{3,l,b}(0) = 0,\ \bar{Z}_{3,l,b}(3) = 4/L,\ \bar{Z}_{3,l,b}(3+4\gamma/L^2) = \bar{Z}_{3,l,b}(6) = 0,\\
&\bar{Z}_{2,a}(0) = \bar{Z}_{2,a}(3) = 0,\ \bar{Z}_{2,a}(3+4\gamma/L^2) = 4-4/L^2,\\
&\quad \bar{Z}_{2,a}(3+4\gamma) = \bar{Z}_{2,a}(6) = 0,\\
&\bar{Z}_{1,a}(0) = \bar{Z}_{1,a}(3) = 0,\ \bar{Z}_{1,a}(3+4\gamma) = 4-16\gamma/3,\ \bar{Z}_{1,a}(6) = 0.
\end{aligned}
\tag{3.2}
$$

One can interpret the behavior of $(\bar{T}(t), \bar{Z}(t))$ as follows, dividing time into the intervals $[0, 3]$ and $[3, 6]$. Initially, only the class $(1, b)$ is occupied. Over times in $[0, 3]$, this mass and that entering the network flow from $(1, b)$ to $(2, b)$, to $(3, l, a)$, $l = 1, \dots, L$, and then to $(3, l, b)$. The mass in $(1, b)$ quickly "drains out" by $t = \gamma/(1-\gamma)$, after which $(1, b)$ continues to process, but at a slower rate. Mass at $(2, b)$ and $(3, l, a)$ continues to be processed at the maximal rate until $t = 3$. The amount of mass at $(2, b)$ rises quickly until $t = \gamma/(1-\gamma)$, after which it gradually drains out until the class is empty at $t = 3$. The classes $(3, l, a)$ remain empty the entire time since $\mu_{2,b} = \sum_{l=1}^{L} \mu_{3,l,a} = 4/3$. Over $[0, 3]$, none of the classes $(3, l, b)$, $(2, a)$ or $(1, a)$ receives service: $(2, a)$ and $(1, a)$ because no mass has gotten that far yet, and $(3, l, b)$ because the classes $(3, l, a)$ are being served at maximal capacity.

At $t = 3$, the behavior of the system changes. Since there is no mass in $(1, b) \cup (2, b) \cup (3, a)$, mass will be processed at the classes $(3, l, b)$, upon which it goes to $(2, a)$. Since this class has priority at the second station, mass there will be processed and will be passed to class $(1, a)$, where processing also immediately begins. The processing of mass at $(1, a)$ and $(2, a)$ immediately causes that at $(1, b)$ and $(2, b)$ to stop. Since $L\mu_{3,l,b} > \mu_{2,a} > \mu_{1,a}$ and there is, at $t = 3$, no mass at $(2, b)$ and $(3, a)$, these classes will remain empty until the mass at $(1, a)$ drains out. This occurs at $t = 6$. [The quicker serving classes $(3, l, b)$ and $(2, a)$ empty at $t = 3 + 4\gamma/L^2$ and $t = 3 + 4\gamma$, respectively.] So, at $t = 6$, the network is empty except at $(1, b)$; over the period $[3, 6]$, three units of mass have accumulated there.

We note that the evolution of $(\bar{T}(t), \bar{Z}(t))$ can also be analyzed for the variant of the network one obtains by formally setting $\gamma = 0$, which corresponds to instantaneous service at the classes $(1, b)$, $(3, l, b)$ and $(2, a)$. As in Lu and Kumar (1991), this type of idealization simplifies the bookkeeping involved. Here, $\bar{Z}(t)$ is piecewise linear over $(0, 3)$ and $(3, 6)$. It is discontinuous at $t = 0$ and $t = 3$, with all of the mass at $(1, b)$ simultaneously flowing to $(2, b)$ (at $t = 0$), and all of the mass at $(3, l, b)$ simultaneously flowing to $(2, a)$, and then to $(1, a)$ (at $t = 3$).

One can check that, over each of the five subintervals of $[0, 6]$ on which $\bar{T}'(t)$ and $\bar{Z}'(t)$ are constant, the pair $(\bar{T}(t), \bar{Z}(t))$, for $\gamma > 0$, solves the fluid model equations (2.10), (2.11). The reasoning is in each case obvious, but takes a little time because of the number of cases involved. We therefore obtain the following proposition.

PROPOSITION 3.1. *The pair* $(\bar{T}(t), \bar{Z}(t))$, *defined in* (3.1), (3.2), *is a solution of the fluid model equations* (2.10), (2.11) *over* $t \in [0, 6]$.

We point out that, in the above construction of $(\bar{T}(t), \bar{Z}(t))$, the behavior at all of the stations $(3, l)$ is identical. By collapsing these stations into a single station 3 as in (2.4), (2.5), it is therefore simple to modify the above pair at $(3, l)$ so that it becomes a fluid model solution for this collapsed network.

By piecing together scaled versions of the above function $(\bar{T}(t), \bar{Z}(t))$, one may extend it to a solution of (2.10), (2.11) over $t \in [0, \infty)$. Specifically, for

$i \in \mathbb{Z}$ and $t \in [3^{i+1}, 3^{i+2}]$, set

$$\widetilde{T}(t) = 3^i\big(\bar{T}(3^{-i}t - 3) + U\big), \qquad \widetilde{Z}(t) = 3^i \bar{Z}(3^{-i}t - 3), \tag{3.3}$$

where U is the constant with components

$$U_{1,b} = U_{2,a} = 2\gamma, \qquad U_{2,b} = U_{3,l,a} = U_{1,a} = \frac{3}{2}, \qquad U_{3,l,b} = \frac{2\gamma}{L^2}, \tag{3.4}$$

and set $\widetilde{T}(0) = \widetilde{Z}(0) = 0$. It is straightforward to check that, over $[3^{i+1}, 3^{i+2}]$, $i \in \mathbb{Z}$, $\widetilde{Z}(t)$ remains a solution of (2.10), (2.11), since the effect of the scaling terms 3^i and 3^{-i} cancel each other out, and since the translation terms are harmless. [$i = 0$ corresponds to $(\bar{T}(t), \bar{Z}(t))$, after a time shift.] By (3.2), $\bar{Z}(6) = 3\bar{Z}(0)$. Therefore, $\widetilde{Z}(t)$ is consistently defined at the endpoints 3^i. One can check that $\widetilde{T}(3^i)$ is also consistently defined by integrating $\widetilde{T}'(t)$ in (3.1) over $[0, 6]$. Faster, though, is to note that, over $[0, 6]$, four units of mass have been processed at each class (lumping $(3, l)$, $l = 1, \dots, L$, together). One then multiplies this by the service times m_k. [Also, note that $U_k = \bar{T}_k(6)/2$ for each k.] One thus obtains that $(\widetilde{T}(t), \widetilde{Z}(t))$ is a solution of (2.10), (2.11) over $(0, \infty)$. Since

$$\lim_{t\to 0} \widetilde{T}(t) = 0, \qquad \lim_{t\to 0} \widetilde{Z}(t) = 0,$$

$(\widetilde{T}(t), \widetilde{Z}(t))$ is continuous at 0, which implies that $(\widetilde{T}(t), \widetilde{Z}(t))$ is, in fact, a solution over $[0, \infty)$, as desired.

It follows from $|\widetilde{Z}(3)| = |\bar{Z}(0)| = 1$ and (3.3), that

$$\limsup_{t\to\infty} |\widetilde{Z}(t)|/t \geq 1/3. \tag{3.5}$$

So $(\widetilde{T}(t), \widetilde{Z}(t))$ is an unstable solution of the fluid model equations (2.10), (2.11). One can, in fact, check, by using (3.2), that the infimum is taken along the sequence 3^i as $i \to \infty$, and so

$$\liminf_{t\to\infty} \big|\widetilde{Z}(t)\big|/t = 1/3. \tag{3.6}$$

Setting $\bar{Z}(t) = \widetilde{Z}(t)$ in (2.12), this implies Theorem 2.

4. Outline of Theorem 1.

We wish to show that the queueing network given in Theorem 1 is stable, that is, its underlying Markov process $Z(t)$ is positive recurrent. The remaining sections of the paper are devoted to this. The purpose of the present section is to rephrase the problem in terms of fluid limits and to summarize the procedure in Sections 5–10.

Fluid limits are a standard tool in showing the stability of queueing networks. Employed in Rybko and Stolyar (1992), they were systematized in Dai (1995) and Stolyar (1994). A *fluid limit* of the queueing network pair $(T(t), Z(t))$ is defined, in our setting, to be any limit

$$\big(\bar{T}(t), \bar{Z}(t)\big) = \lim_{n\to\infty} \frac{1}{|z_n|}\big(T^{z_n}(t|z_n|), Z^{z_n}(t|z_n|)\big), \tag{4.1}$$

for any sequence z_n with $|z_n| \to \infty$, and any ω, which satisfies the fluid model equations (2.10), (2.11). [$Z(0) = z_n$; as before, $|\cdot|$ denotes the sum norm.] Convergence is required to be uniform on compact sets of t (u.o.c.). One automatically has $|\bar{Z}(0)| = 1$. The *fluid limit model* is said to be *stable* when all fluid limits satisfy $\bar{Z}(t) = 0$ for $t \geq \delta$, for some $\delta > 0$.

Here, we employ a somewhat weaker version of fluid limits, which is more natural in our setting. We define a *fluid limit on H^z*, for events H^z with $\lim_{|z|\to\infty} P(H^z) = 1$, to be any fluid limit in (4.1), for which $\omega \in H^{z_n}$ for all n. The fluid limit model is *asymptotically stable* if there exist such H^z, so that the condition $\bar{Z}(t) = 0$ for $t \geq \delta$, $\delta > 0$, holds for all fluid limits on H^z. This modification adds a degree of flexibility to the use of fluid limits. In our case, it will allow us to exclude the "bad" events where $Z_3(t)/t$ does not remain small, after an initial adjustment period.

In order to employ the limits in (4.1), we need to know that they are fluid limits (i.e., that they satisfy the fluid model equations) on a set of probability 1. There is a standard framework for this. Let G denote the event where the strong law of large numbers holds, in each case, for the sums of the interarrival times, the service times at each class, and the routing vectors of the network. The interarrival and service times consist of independent exponential random variables; the routing vectors are deterministic except following departures from class $(2, b)$, at which point each of the L classes $(3, l, a)$ is chosen with equal probability, independently of past choices. So, the strong law holds almost surely for the corresponding sums, that is, $P(G) = 1$. Consider now $(T^{z_n}(t), Z^{z_n}(t))$ along any sequence z_n, with $|z_n| \to \infty$ as $n \to \infty$. We will require that, on G, each such sequence possess a subsequence z_{i_n} (depending on ω) on which convergence is u.o.c. and where

$$(4.2) \qquad \lim_{n\to\infty} \frac{1}{|z_{i_n}|}\big(T^{z_{i_n}}(t|z_{i_n}|), Z^{z_{i_n}}(t|z_{i_n}|)\big) \quad \text{is a fluid limit.}$$

As mentioned at the beginning of the section, we wish to rephrase Theorem 1 in terms of fluid limits. The following result enables us to do this. It is a special case of Theorem 3′ in Bramson (1998), which is a modification of Theorem 4.3 in Dai (1995). (In Theorem 4.3, the fluid limit model is assumed to be stable; in Theorem 3′, it is assumed to be asymptotically stable.)

THEOREM 3. *Suppose that for a given L,* (i) *the fluid limit model corresponding to the queueing network in* (2.1), (2.2) *is asymptotically stable and that* (ii) (4.2) *holds. Then, the queueing network is stable.*

Once we verify conditions (i) and (ii) for the queueing networks in (2.1), (2.2) and large enough L, Theorem 1 will follow immediately from Theorem 3. As we will explain shortly, all of the work is concentrated in checking condition (i).

We point out here that the conclusion in Theorem 3 can be strengthened. Namely, under (i) and (ii), the equilibrium distribution of the queueing network, in fact, has all positive moments. This would follow immediately from Theorem 4.1 of Dai and Meyn (1995), if in place of (i), one had the assumption that (i′) the fluid model corresponding to the network in (2.1), (2.2) is stable. One can check, though, that the proof of Theorem 4.1 follows as before, if one instead assumes (i), because either version suffices for Proposition 5.1 in Dai and Meyn (1995). The remaining parts of the argument are more general and build on the proposition (personal communication from J. Dai).

The verification of condition (ii) in Theorem 1, for the queueing network in (2.1), (2.2) and any L, is fairly standard. It involves repeated application of the strong law of large numbers, and use of the uniform convergence of $(1/|z_{i_n}|)T^{z_{i_n}}(t|z_{i_n}|)$ and $(1/|z_{i_n}|)Z^{z_{i_n}}(t|z_{i_n}|)$ over bounded t. Condition (ii), for related networks, is shown in the proof of Theorem 4.1 in Dai (1995) and is summarized below Theorem 5 in Bramson (1998). The present setting is actually somewhat simpler; we are able to omit residual times here, since all of the times are exponentially distributed.

We now summarize the reasoning we employ to verify condition (i) of Theorem 3. The reasoning involves a number of steps and occupies the remainder of the paper. The goal is to show that, for large enough L, the fluid limit model corresponding to the network in (2.1), (2.2) is asymptotically stable. As was shown in Section 3, the fluid model corresponding to the network is not stable. The difference in behavior for the two cases is due to the fluctuations which are present for the queueing network, but not for the fluid model. In order to understand the fluid limit model, we analyze their effect.

The crucial behavior in the queueing network occurs at the classes $(3, l, a)$, $l = 1, \dots, L$. The main point here is that, for large L, it is much more likely for $Z_{3,a}(t)$ to decrease by a factor of 2 rather than increase by this factor, when $Z_{3,a}(t)$ is large. Appropriate bounds for this are given in Propositions 6.1 and 6.3. The basic idea is as follows. One chooses L to be large in order to increase the probability that some (random) class $(3, l_0, a)$ becomes empty before $Z_{3,a}(t)$ doubles. Once the former occurs, service begins at $(3, l_0, b)$. The customers who have been served at $(3, l_0, b)$ immediately begin service at the high priority class $(2, a)$. This, in turn, prevents customers at $(2, b)$ from being served until $(2, a)$ is empty. Without this interference from $(2, a)$, the service rate at $(2, b)$ is the same as the combined service rate at $(3, l, a)$, $l = 1, \dots, L$, which is 4/3. This interference, however, creates idle periods for $(2, b)$, which, in effect, cause it to have a slower service rate than at $(3, a)$. As a consequence, $Z_{3,a}(t)$ will typically decrease by a fixed factor, say 2, before it increases by the same factor. (In fact, for each l, $Z_{3,l,a}(t) = 0$ will likely occur over this time span, although not necessarily at a common time.)

In Section 5, basic large deviation bounds are given which apply in this setting. Together with Propositions 6.1 and 6.3, they imply Proposition 6.4, which states that, under “small” $Z_{3,a}(0)$, $Z_{3,a}(t)$ will typically remain comparatively small for an extended period of time. Since the proof of Proposition 6.3 is rather long, it is carried out in Section 7.

In Section 8, it is shown that, under “small” $Z_{3,a}(0)$ and $Z_{3,l,b}(0)$, for a given l, $Z_{3,l,b}(t)$ will also typically remain small for an extended period of time. The main result here is Proposition 8.2. It relies on bounds on the incremental growth of $Z_{3,l,b}(t)$, for given l, from Proposition 8.1 and its corollary, and on the bounds from Section 5. Less work is required here than for Proposition 6.4, since one already has upper bounds on the size of $Z_{3,a}(t)$ over this time span.

For the bounds on $Z_{3,a}(t)$ and $Z_{3,l,b}(t)$ in Sections 6 and 8, $Z_{3,a}(0)$ and $Z_{3,l,b}(0)$ were assumed to be small. To be able to apply these bounds for general initial data, one needs to be able to restart $Z(t)$ where these values are small. It is shown in Section 9 that such stopping times exist. Work is required to show this for $Z_{3,a}(t)$, since similar behavior is required simultaneously at all classes $(3, l, a)$. The basic point is that when customers at $(2, b)$ are not being served, this is felt simultaneously at $(3, l, a)$ for all l. It is considerably easier to analyze $Z_{3,l,b}(t)$, since we only require a bound for a given l at a specific time, and since $(3, l)$ is strictly subcritical. These results for $Z_{3,a}(t)$ and $Z_{3,l,b}(t)$ are given in Propositions 9.3 and 9.4.

In Section 10, we demonstrate (i) of Theorem 3. Employing Propositions 6.4, 8.2, 9.3 and 9.4, we show, in Proposition 10.1, that for appropriate sets H^z, with $P(H^z) \to 1$ as $|z| \to \infty$, all fluid limits $\bar{Z}(t)$ on H^z satisfy $\bar{Z}_3(t) = 0$ for $t \geq t_0$ and appropriate t_0. One can therefore, in effect, omit station 3 from the system when analyzing the behavior of the fluid limits. This reduces the system to the two-station, four-class network in (2.6), (2.7), where the discipline is last-buffer-first-served. Such networks are known to be stable [Dai and Weiss (1996)]. This enables us to show, in Proposition 10.3, that, for large L, our fluid limit model is asymptotically stable, and so (i) of Theorem 3 holds.

5. Large deviation bounds. Here, we present some large deviation bounds that will be used in Sections 6–9. The main results, Proposition 5.1 and its corollary, state that the value of a Markov process will typically remain small for an extended period of time if the process satisfies appropriate negative drift conditions.

First, we recall several elementary large deviation estimates involving exponential and Bernoulli random variables, and symmetric random walks. Let $Y_1, Y_2, \ldots$ be i.i.d. mean-1 exponential random variables, with $S_n = Y_1 + \cdots + Y_n$. Then, for each $\theta > 0$, there exists a $c > 0$, so that for large n,

$$P\left(\frac{1}{n}|S_n - n| \geq \theta\right) \leq e^{-cn}. \tag{5.1}$$

The bound (5.1) can be demonstrated in the usual way by applying Markov's inequality to the moment generating functions of S_n. Note that (5.1) immediately extends to exponential distributions with other means, after scaling. Corresponding large deviation bounds for Poisson random variables also follow from (5.1).

If Y_1, $Y_2, \ldots$ are replaced by i.i.d. mean-m Bernoulli random variables, then the analog for binomial distributions,

$$P\left(\frac{1}{n}|S_n - mn| \geq \theta\right) \leq e^{-cn}, \tag{5.2}$$

holds for each $\theta > 0$ and m, and appropriate $c > 0$. By adding up the exceptional probabilities, (5.2) also gives an upper bound on the probability that any of the components of a normalized multinomial distribution differs by more than θ from its mean. Let X_t denote a rate-1 symmetric nearest neighbor random walk on $\mathbb{Z}$, with $X_0 = 0$. Application of the moment generating functions of X_t, together with the reflection principle, also implies that for each $\beta \in (1/2, 1]$ and $\theta > 0$,

$$(5.3) \qquad P\Bigl(\sup_{s\le t} X_s \ge \theta t^{\beta}\Bigr) \le \exp\{-ct^{2\beta-1}\}$$

for appropriate $c > 0$ and large t. The bound in (5.1) will be applied to the cumulative service times required by customers in specific classes of our queueing network over given stretches of time. The bound in (5.2) will be applied to the proportion of time the different classes $(3, l, a)$, $l = 1, \dots, L$, are selected by customers departing from $(2, b)$. We will apply (5.3) in Proposition 9.1.

Let $W(n)$, $n = 0, 1, 2, \dots$, be a Markov chain with countable state space $\mathscr{I}$. For $i \in \mathscr{I}$, $\|i\|$ denotes a map into $[0, \infty)$. Set $\mathscr{I}_\nu = \{i\colon \|i\| \ge \nu\}$. Let $h_{i,i'} = \|i'\| - \|i\|$, for $i, i' \in \mathscr{I}$, and let $p_{i,i'}$ be the transition probability of $W(n)$ from i to i'. We assume that the upward jump size is uniformly bounded, with

$$(5.4) \qquad p_{i,i'} = 0 \quad \text{for } h_{i,i'} > Q_1,$$

for appropriate Q_1. We also assume that for given $Q_2 > 0$ and $\eta > 0$,

$$(5.5) \qquad \sum_{\mathscr{A}} p_{i,i'} \le \eta \quad \text{for all } i \in \mathscr{I}_\nu,$$

where $\mathscr{A} = \{i'\colon h_{i,i'} > -Q_2\}$. The latter condition says that if η is small, $W(n)$ decreases most of the time by at least Q_2, when $\|W(n)\| \ge \nu$. [For small η, conditions (5.4) and (5.5) can be thought of as a strong variant of Foster's criterion.]

In Lemma 5.1, we give upper bounds on $\|W(n)\|$, when $\eta > 0$ is sufficiently small, for fixed Q_1 and Q_2. The estimates follow by applying Chebyshev's inequality to the moment generating functions of $W(n)$.

LEMMA 5.1. *Assume that the Markov chain $W(n)$ satisfies* (5.4) *and* (5.5) *for given Q_1, Q_2 and ν, and that $\|W(0)\| \le N$, with $N \ge \nu$. For fixed $r > 0$ and small enough $\eta > 0$* (*depending on Q_1, Q_2 and r*),

$$(5.6) \qquad P\bigl(\|W(n)\| \ge M + N \textit{ for some } n \le e^{Mr}\bigr) \le e^{-Mr}$$

holds for large enough M (*depending on Q_1*).

PROOF. Let $U(n) = \exp\{r(\|W(n)\| - N)\} - cn$, where $c = \exp\{Q_1 r\} > 0$. For small enough $\eta > 0$ (depending on Q_1, Q_2 and r), $\{U(n)\}$ is a supermartingale with respect to the σ-algebra $\mathscr{F}_n$ generated by $\{W(1), \dots, W(n)\}$. To see this, note that for $W(n) = i$,

$$(5.7) \quad E\bigl[U(n+1) \mid \mathscr{F}_n\bigr] - U(n) = \exp\bigl\{r(\|i\| - N)\bigr\}\Bigl(\sum_{i'} p_{i,i'} \exp\bigl\{r h_{i,i'}\bigr\} - 1\Bigr) - c.$$

By (5.4) and (5.5), for all $i \in \mathscr{I}_\nu$,

$$\sum_{i'} p_{i,i'} \exp\{rh_{i,i'}\} - 1 \le \eta \exp\{Q_1 r\} + (1-\eta)\exp\{-Q_2 r\} - 1. \tag{5.8}$$

This is negative for small enough η, and so the same holds for (5.7). For $i \notin \mathscr{I}_\nu$, $\exp\{r(\|i\| - N)\} < 1$, and so, because of the choice of c, (5.7) is negative in this case as well. Therefore, $\{U(n)\}$ is a supermartingale.

Let $T = \min\{n\colon \|W(n)\| \ge M + N\} \wedge \lfloor e^{Mr}/2c \rfloor$, for a given M, where $\lfloor x \rfloor$ denotes the integer part of x. Then, by the optional sampling theorem,

$$E[U(T)] \le U(0) \le 1. \tag{5.9}$$

On the event where $\|W(T)\| \ge M + N$, one has $U(T) \ge e^{Mr}/2$. So, by (5.9) and Chebyshev's inequality,

$$P\big(\|W(n)\| \ge M + N \text{ for some } n \le e^{Mr}/2c\big) \le 2e^{-Mr}$$

holds for all M. By decreasing r by 1 so as to absorb the coefficients, we obtain (5.6). □

We will apply Lemma 5.1 in the following setting. Let $Y(t)$, $t \ge 0$, be a Markov process on a countable state space $\mathscr{I}$. Define $\|\cdot\|$ and $\mathscr{I}_\nu$ as before. Assume that $Y(t)$ has jump rates that are bounded above by Γ, $\Gamma > 0$, and that the size of any upward jump is bounded above by J. For given R_1 and R_2, with $R_1, R_2 > 0$, let $q_i(R_1, R_2)$ denote the probability that, starting from i, $\|Y(t)\| - \|i\|$ first exits the interval $(-R_2, R_1)$ on the right. The amount R_1 is "overshot" upon exiting on the right is bounded by J, which will later be assumed to be small relative to R_1. Also, we assume that

$$q_i(R_1, R_2) \le \eta \quad \text{for all } i \in \mathscr{I}_\nu, \tag{5.10}$$

for given ν and η.

Proposition 5.1 gives upper bounds on $\|Y(t)\|$. It will be employed in Proposition 8.2.

PROPOSITION 5.1. *Assume that the Markov process $Y(t)$ is chosen as above, for given Γ, ν, J, R_1 and R_2, and that $\|Y(0)\| \le N$, with $N \ge \nu$. For fixed $r > 0$ and small enough $\eta > 0$ (depending on J, R_1, R_2 and r),*

$$P\big(\|Y(t)\| \ge M + N \text{ for some } t \le e^{Mr}\big) \le e^{-Mr} \tag{5.11}$$

holds for large enough M (depending on Γ, J and R_1).

PROOF. Let $W(n)$ denote the embedded Markov chain on $\mathscr{I}$ formed by stopping $Y(t)$ each time $\|Y(t)\|$ increases by at least R_1 or decreases by at least R_2 from the previous stopped state, $W(n-1)$. Let $\tau(n)$ denote the random time at which the nth such stopped state occurs. The size of each upward jump of $\|W(n)\|$ is at most $R_1 + J$ and that of each downward jump is at least R_2. Setting $Q_1 = R_1 + J$ and $Q_2 = R_2$, it follows from (5.6) that

$$\begin{aligned} &P\big(\|Y(\tau(n))\| \geq M + N \text{ for some } n \leq e^{Mr}\big) \\ &\qquad = P\big(\|W(n)\| \geq M + N \text{ for some } n \leq e^{Mr}\big) \leq e^{-Mr} \end{aligned} \tag{5.12}$$

holds for large enough M, if $\|W(0)\| \leq N$, with $N \geq \nu$.

It follows from (5.1) that, for appropriate $c > 0$ and large n,

$$P\big(\tau(n) \leq n/2\Gamma\big) \leq e^{-cn}.$$

Together with (5.12), this implies that

$$P\big(\|Y(t)\| \geq M + N \text{ for some } t \leq \exp\{Mr\}/2\Gamma\big) \leq \exp\{-Mr\} + \exp\{-ce^{Mr}\}$$

for large M. Decreasing r by 1 so as to absorb Γ implies (5.11). □

We will also employ the following variant of the process $Y(t)$ in Proposition 5.1. As before, $Y(t)$ is a Markov process on a countable state space $\mathscr{S}$, with jump rates bounded above by $\Gamma > 0$, and the size of upward jumps bounded above by J. For given $R_1, R_2 > 1$, we now let $q_i(R_1, R_2)$ denote the probability, starting from i, with $\|i\| \geq 1$, that $\|Y(t)\|/\|i\|$ first exits the interval $(1/R_2, R_1)$ on the right. As before, we assume that (5.10) holds, but for this new choice of $q_i(R_1, R_2)$ and with $\nu \geq 1$.

Setting $\|i\|' = (\log \|i\|) \vee 0$, the conditions of Proposition 5.1 are satisfied, for new choices of ν, J, R_1 and R_2. We therefore obtain the following multiplicative variant of Proposition 5.1. It will be employed in Proposition 6.4.

COROLLARY 5.1. *Assume that the Markov process $Y(t)$ is chosen as above, for given Γ, ν, J, R_1 and R_2, and that $\|Y(0)\| \leq N$, with $N \geq \nu$. For fixed $r > 0$ and small enough $\eta > 0$ (depending on J, R_1, R_2 and r),*

$$P\big(\|Y(t)\| \geq MN \textit{ for some } t \leq M^r\big) \leq M^{-r} \tag{5.13}$$

holds for large enough M (depending on Γ, J and R_1).

6. Bounds on the classes (3, *a*). In this section, we obtain upper bounds on the total number of customers at $(3, a)$, the union of the classes $(3, l, a)$, $l = 1, \ldots, L$. We show, in Proposition 6.4, that if the number of customers at $(3, a)$ is initially small on the appropriate scale and L is large, then the total number of customers at $(3, a)$ will remain comparatively small for a long period of time. These estimates are employed in Section 10 to show that, for the corresponding fluid limits, the amount of mass at $(3, a)$ remains 0 for all time, for fluid limits starting with no mass there. These estimates are also needed, in Section 8, to obtain upper bounds on the number of customers at $(3, b)$.

Our approach in deriving Proposition 6.4 will be to show that, for large L, the number of customers at $(3, a)$, $Z_{3,a}(t)$, satisfies the assumptions of the corollary to Proposition 5.1. This involves showing that when $Z_{3,a}(t)$ is large enough, $Z_{3,a}(t)$ will typically decrease by a factor of 2 before it increases by this factor. Propositions 6.1 and 6.3 contain the bounds required for this.

Proposition 6.1 gives upper bounds on the probability of $Z_{3,a}(t)$ doubling by an appropriate time, together with partial bounds on it decreasing by a factor of 2 by then. In order to complete the latter direction, one needs to show that each class $(3, l, a)$, $l = 1, \dots, L$, will, with high probability, be empty at some time before then. Proposition 6.3 supplies this bound. One of the main steps is given in Proposition 6.2, which deals with the likelihood of a class $(3, l, b)$, for a chosen l, continuing to serve all of its customers once it has started, without being interrupted by the class $(3, l, a)$. The proof of Proposition 6.3 itself is rather long; we only summarize it here, deferring the proof to Section 7.

In order to derive Proposition 6.1, we first show Lemmas 6.1 and 6.2. Lemma 6.1 gives elementary upper bounds on the probability of an L-dimensional Brownian motion doubling in size by an appropriate time, and lower bounds on the probability of individual coordinates hitting 0. Lemma 6.2 modifies these bounds to symmetric random walks. In Proposition 6.1, the bounds on these symmetric random walks are applied to $Z_{3,a}(t)$.

Lemma 6.1 is an elementary result on L-dimensional Brownian motion $B(s) = (B_1(s), \dots, B_L(s))$, whose components have variances s/L^2 and no drift. For $s_1 > 0$, we let $H_B(s_1)$ denote the event on which at least one component $B_l(s)$ hits 0 on $[s_1/4, s_1/2]$. As always, $|\cdot|$ denotes the sum norm.

LEMMA 6.1. *Assume that* $|B(0)| \le 1$. *For a given* $\eta > 0$, *and* $s_1 > 0$ *sufficiently small, not depending on* L,

$$P\Big(\sup_{s \le s_1} |B(s)| \ge 2\Big) \le \eta. \tag{6.1}$$

For large enough L,

$$P\big(H_B^c(s_1)\big) \le \eta. \tag{6.2}$$

PROOF. For each l, $l = 1, \dots, L$, $|B_l(s_1) - B_l(0)|$ has mean $\sqrt{2s_1/\pi}/L$. So, $|B(s_1) - B(0)|$ has mean $\sqrt{2s_1/\pi}$, and, by Chebyshev's inequality,

$$P\big(|B(s_1) - B(0)| \ge 1\big) \le \eta/2$$

for $s_1 \le \pi\eta^2/8$. Stop the process $B(s) - B(0)$ as soon as $|B(s) - B(0)| = 1$. By reflecting $B(s) - B(0)$ across the corresponding face of the diamond $|x - B(0)| = 1$, it follows that the probability that $|B(s_1) - B(0)| \ge 1$ is at least one half the probability that $|B(s) - B(0)| = 1$ at some $s \le s_1$. Consequently,

$$P\Big(\sup_{s \le s_1} |B(s) - B(0)| \ge 1\Big) \le \eta.$$

This implies (6.1).

Note that $|B_l(0)| \le 2/L$ must hold for at least half of the indices. For each such index, the probability of the corresponding Brownian motion hitting 0 over $[s_1/4, s_1/2]$ is the same as standard Brownian motion, with initial value in $[-2, 2]$, hitting 0 over the same interval, which is bounded away from 0. Since $B_l(s)$, $l = 1, \dots, L$, are independent, (6.2) follows immediately, for fixed $s_1 > 0$ and large enough L. □

Let $R_1(t), \dots, R_L(t)$ be a family of i.i.d. symmetric nearest neighbor continuous time random walks on $\mathbb{Z}$, each with jump rate $1/L^2$, and let $R(t)$ denote the corresponding vector. For $t_1 > 0$, let $H_R(t_1)$ denote the event on which at least one component $R_l(t)$ hits 0 over $[t_1/4, t_1/2]$. Lemma 6.2 is the analog of Lemma 6.1, but with these random walks replacing the above Brownian motions. Here and later on in Sections 6 and 7, s_1 will be small but fixed (after the choice of η), L large but fixed (after the choice of s_1), and N will be sufficiently large.

LEMMA 6.2. *For a given $\eta > 0$, let $s_1 > 0$ be sufficiently small and L (depending on s_1) be sufficiently large. Assume that $|R(0)| \le N$, where N is sufficiently large. Then,*

$$P\Big(\sup_{t \le N^2 s_1} |R(t)| \ge 2N\Big) \le \eta \tag{6.3}$$

and

$$P\big(H_R^c(N^2 s_1)\big) \le \eta. \tag{6.4}$$

We summarize the reasoning behind the lemma. It follows from a standard form of the invariance principle that $R(N^2 s)/N$ converges in distribution, in the Skorokhod topology on $[0, s_1]$, to $B(s)$, if the initial data converge [see, e.g., Billingsley (1968) or Ethier and Kurtz (1986)]. The set of paths in (6.1) is closed. So, (6.3) follows from (6.1). The time for a Brownian motion to hit 0 is a.s. continuous and has probability 0 of occurring at any specific time. One can therefore check that (6.4) follows from (6.2).

We can apply Lemma 6.2 to the number of customers $Z_{3,l,a}(t)$ in the classes $(3, l, a)$, $l = 1, \dots, L$, by means of an elementary comparison. Recall that customers in the classes $(3, l, a)$ are served at rate $4/3L$ for each l. Also, recall that customers, in the class $(2, b)$ leading to $(3, a)$, are served at rate $4/3$, after which a customer chooses one of the L classes $(3, l, a)$ with equal probability. Either because $(2, b)$ is empty or the higher priority class $(2, a)$ is not, service at $(2, b)$ may be suppressed. Using “ghost” service times in these instances, it is easy to couple the L-tuple $(Z_{3,1,a}(t), \dots, Z_{3,L,a}(t))$ pathwise to $(R_1(t), \dots, R_L(t))$ so that, for all l and t,

$$Z_{3,l,a}(t) \le \left| R_l\Big(\frac{8}{3} L t\Big) \right|, \quad \text{with } Z_{3,l,a}(0) = R_l(0), \tag{6.5}$$

always holds. Denote by $H_{3,a}(t_1)$ the event on which at least one component $Z_{3,l,a}(t)$ hits 0 over $[t_1/4, t_1/2]$, and by τ_a the first time by which all of the components have hit 0, that is,

$$\tau_a = \max_l \inf\big\{t\colon Z_{3,l,a}(t) = 0\big\}.$$

In Proposition 6.1, the bounds (6.6) and (6.8) correspond to (6.3) and (6.4) in Lemma 6.2, and the bound (6.7) is a modification of (6.3). For the remainder of this section and in Section 7, we will set $t_1 = N^2 s_1/L$.

PROPOSITION 6.1. *For a given $\eta > 0$, let $s_1 > 0$ be sufficiently small and L (depending on s_1) be sufficiently large. Assume that $Z_{3,a}(0) \le N$, where N is sufficiently large. Then,*

$$P\Bigl(\sup_{t\le t_1} Z_{3,a}(t) \ge 2N\Bigr) \le \eta, \tag{6.6}$$

$$P\Bigl(\sup_{t\in[\tau_a,\, t_1]} Z_{3,a}(t) \ge N/2\Bigr) \le \eta \tag{6.7}$$

and

$$P(H^c_{3,a}(t_1)) \le \eta. \tag{6.8}$$

PROOF. The bounds in (6.6) and (6.8) follow immediately from (6.3) and (6.4), and the comparison in (6.5), after reducing s_1 by the factor 8/3. In order to obtain (6.7), let $\widetilde{Z}(t)$ denote the process obtained from $\widehat{Z}(t) = (Z_{3,1,a}(t), \dots, Z_{3,L,a}(t))$ by redefining the motion in each coordinate to be a symmetric nearest neighbor random walk with jump rate $8/3L$ after that coordinate first hits 0. One can couple the two processes $\widetilde{Z}(t)$ and $\widehat{Z}(t)$ so that

$$Z_{3,a}(t) = \bigl|\widehat{Z}(t)\bigr| \le |\widetilde{Z}(t)|. \tag{6.9}$$

Let $\widetilde{R}(t) = R((8/3)Lt)$ denote the scaled process of random walks with $\widetilde{R}(0) = R(0) = 0$ and rate $8/3L$ in each coordinate. Since $\widetilde{Z}_l(t) = 0 \le |\widetilde{R}_l(t)|$ when $\widetilde{Z}_l(t)$ hits 0, one can also couple these processes so that

$$\bigl|\widetilde{Z}(t)\bigr| \le \bigl|\widetilde{R}(t)\bigr| \quad \text{for } t \ge \tau_a. \tag{6.10}$$

The bound (6.7) then follows from (6.3) and (6.9), (6.10), after increasing N by a factor of 4 and decreasing s_1 by a factor of $32/3$. □

In Proposition 6.1, (6.6) provides an upper bound on the probability of $Z_{3,a}(t)$ doubling by time t_1. If we knew that $\tau_a \le t_1$ typically holds, then (6.7) would show that $Z_{3,a}(t_1) < N/2$ also typically holds. By (6.8), we know that at least one class $(3, l, a)$ will typically be empty by then; for $\tau_a \le t_1$, we need to show that this is true for all L such classes. This result is shown in Proposition 6.3. The basic idea is that once customers from a class $(3, l, b)$ are being served, the class $(2, a)$ will receive a supply of customers, which prevents service at $(2, b)$. This, in turn, gives all of the other classes $(3, l, a)$ the chance to become empty, and hence for τ_a to occur.

As a first step in this direction, we demonstrate Proposition 6.2. The result states, in essence, that once service begins at some class $(3, l, b)$, customers are prevented from entering the corresponding class $(3, l, a)$ from $(2, b)$ until $(3, l, b)$ is empty. More precisely, let σ denote the first time t, $t \ge t_1/4$, at which

$Z_{3,l,a}(t) = 0$ holds for some l, which we denote by l_0. We write F_1 for the event where (i) $\sigma \le t_1/2$ and (ii) the first time after σ, at which $Z_{3,l_0,a}(t) > 0$, is greater than the first time after σ, at which $Z_{3,l_0,b}(t) = 0$. The basic idea of the proof is to show that, off of a set of small probability [because of the choice of $m_{3,l,b}$ in (2.2)], once σ occurs, $Z_{2,a}(t)$ dominates a birth–death process until $Z_{3,l_0,b}(t) = 0$. A related comparison is used in Proposition 8.1.

PROPOSITION 6.2. *For a given $\eta > 0$, let $s_1 > 0$ be sufficiently small, and L (depending on s_1) be sufficiently large. Assume that $Z_{3,a}(0) \le N$, where N is sufficiently large. Then,*

$$P(F_1^c) \le \eta. \tag{6.11}$$

PROOF. We first observe that $Z_{2,a}(t)$ can be compared to a birth–death process on $\{0, 1, 2, \dots\}$, with birth rate L/γ and death rate $1/\gamma$, and both processes starting from the same initial state. To see this, recall that customers in the class $(2, a)$ are served at rate $1/\gamma$. Also, recall that customers in each class $(3, l, b)$ are served at rate L/γ, after which they enter the class $(2, a)$, with service being continued at a class $(3, l, b)$ as long as (a) $(3, l, a)$ is empty and (b) $(3, l, b)$ is not. So, $Z_{2,a}(t)$ dominates a copy of the above birth–death process up until the time either (a) or (b) fails for a specified l. Moreover, no customer in $(2, b)$ can be served as long as (a′) $Z_{2,a}(t) > 0$, in which case $(3, l, a)$ remains empty. So, condition (a′) can be substituted for (a) if $(3, l, a)$ is initially empty.

Assume now that the event $H_{3,a}(t_1)$ occurs, and that σ and l_0 are defined as above. We recall that, by (6.8), $H_{3,a}(t_1)$ occurs off of a set of probability η. On $H_{3,a}(t_1)$, one automatically has $\sigma \le t_1/2$. Now, restart $Z(t)$ at time σ. By the strong Markov property, the restarted Markov process evolves according to the same law as before. We claim that, irrespective of σ and $Z(\sigma)$, the class $(3, l_0, b)$ will be empty before $(3, l_0, a)$ is occupied again, off of a set of probability η. This will imply the desired inequality (6.11), for a new choice of η which is twice the original value.

If $Z_{3,l_0,b}(\sigma) = 0$, then we are already done. If $Z_{3,l_0,b}(\sigma) > 0$, then, off of a set of probability at most $4\gamma/3L^2 \le 1/L$, the class $(2, a)$ will be occupied before $(3, l_0, a)$ is. Assume that this nonexceptional event occurs, with σ', $\sigma' \ge \sigma$, denoting the corresponding time. Restart the process at time σ'. As shown in the first paragraph of the proof, $Z_{2,a}(t+\sigma')$ dominates a birth–death process with rates L/γ and $1/\gamma$ until either (a′) or (b) fails. Here, the birth–death process has initial value at least 1. It is therefore routine to check that the probability of this birth–death process ever reaching 0 is at most $1/L$. By the above domination, the probability of the class $(2, a)$ becoming empty [and thus $(3, l_0, a)$ possibly being occupied], before (b) fails, is therefore at most $1/L$. Adding this to the exceptional probability at the beginning of the paragraph shows that, off of a set of probability $2/L$, starting at time σ, the class $(3, l_0, b)$ will be empty before $(3, l_0, a)$ is occupied again. Choosing $L \ge 2/\eta$ gives the desired bound η, and hence completes the proof of (6.11). □

Proposition 6.3 gives an upper bound on the time required for all classes $(3, l, a)$ to empty, in terms of $Z_{3,a}(0)$. Recall that $t_1 = N^2 s_1/L$.

PROPOSITION 6.3. *For a given* $\eta > 0$, *let* $s_1 > 0$ *be sufficiently small, and* L (*depending on* s_1) *be sufficiently large. Assume that* $Z_{3,a}(0) \le N$, *where* N *is sufficiently large. Then,*

$$P(\tau_a > t_1) \le \eta. \tag{6.12}$$

Together with (6.7) of Proposition 6.1, Proposition 6.3 implies the following corollary, which provides lower bounds on the probability of $Z_{3,a}(t)$ decreasing by a factor of 2. The bounds in (6.6) and (6.13) are the main estimates required for Proposition 6.4.

COROLLARY 6.1. *For a given* $\eta > 0$, *let* $s_1 > 0$ *be sufficiently small and* L (*depending on* s_1) *be sufficiently large. Assume that* $Z_{3,a}(0) \le N$, *where* N *is sufficiently large. Then,*

$$P\big(Z_{3,a}(t_1) \ge N/2\big) \le \eta. \tag{6.13}$$

The proof of Proposition 6.3 is fairly long, and so we postpone it until the next section. We note here that the argument breaks into two cases, depending on whether the total number of customers to visit $(3, a)$ $(= \bigcup_{l=1}^{L}(3, l, a))$ by time $t_1/4$ is at most $t_1/5L$ or is greater than this. In the first case, the number of customers is small enough to directly show, in Proposition 7.1, that $\tau_a \le t_1/4$ typically holds. The second case is more complicated and is handled in Proposition 7.2. There, it is shown that the number of customers served at class $(2, a)$, by time t_1, is typically large enough to substantially slow down the influx of customers from class $(2, b)$ to $(3, a)$, which will imply that $\tau_a \le t_1$ typically holds.

Proposition 6.4 is the main result of this section. It states that if the number of customers at $(3, a)$ is initially at most a given power of M, then this will typically continue to be the case, for any larger power, for a long period of time relative to M.

PROPOSITION 6.4. *For given* $0 < r_0 < r_1$, *let* L *be sufficiently large. Assume that* $Z_{3,a}(0) \le M^{r_0}$, *where* M *is sufficiently large. Then,*

$$P\big(Z_{3,a}(t) \ge M^{r_1} \text{ for some } t \le M^2\big) \le 1/M. \tag{6.14}$$

For our purposes, $0 < r_0 < r_1 < 1/2$ is the relevant range of r_0 and r_1. When analyzing the fluid limits of $Z(t)$ in Section 10, we will scale time and $Z(t)$ by M; this, in particular, implies that the upper bound on time in (6.14) goes to ∞ as $M \to \infty$. Although we do not require this here, the bounds M^2 and $1/M$ in (6.14) can easily be replaced by any power.

PROOF OF PROPOSITION 6.4. Define $\|\cdot\|$ so that $\|Z(t)\| = Z_{3,a}(t)$. It is not difficult to verify that the conditions of Corollary 5.1 are satisfied for $Z(t)$ and

$\|\cdot\|$. Recall that $Z(t)$ is a Markov process with jump rates at most $\Gamma = 6L^2/\gamma$, and upward jumps at most $J = 1$. For given $Z(0)$, with $Z_{3,a}(0) = N$, consider the probability that $Z_{3,a}(t)/N$ first exits the interval $(1/2, 2)$ on the right. By (6.6) and (6.13), for sufficiently large L and N, this probability can be chosen as close to 0 as desired. So, for given $\eta > 0$ and large L, (5.10) is satisfied by $Z(t)$ and $\|\cdot\|$, with $R_1 = R_2 = 2$ and appropriate ν.

It follows from (5.13) of Corollary 5.1, that for fixed r and large enough L,

$$P\big(Z_{3,a}(t) \geq MN \text{ for some } t \leq M^r\big) \leq M^{-r} \tag{6.15}$$

holds for $Z_{3,a}(0) \leq N$ and large enough M and N. Substitution of $M^{r_1-r_0}$ for M, M^{r_0} for N and $2/(r_1 - r_0)$ for r, in (6.15), shows that

$$P\big(Z_{3,a}(t) \geq M^{r_1} \text{ for some } t \leq M^2\big) \leq 1/M$$

holds for $Z_{3,a}(0) \leq M^{r_0}$ and large enough M, which is (6.14). □

7. Proof of Proposition 6.3.

In this section, we demonstrate Proposition 6.3. We break the bound there, (6.12), into three parts. One part, a bound on $P(F_1^c)$, is already given in Proposition 6.2. The other two parts are treated in Propositions 7.1 and 7.2. To state them, we let F_2 denote the event such that, by time $t_1/4$, the total number of customers to visit $(3, a)$ is strictly greater than $t_1/5L$. [This quantity includes customers originally at $(3, a)$.] The assertions (7.1) and (7.2) in Theorems 7.1 and 7.2, respectively, are then adapted from (6.12) by restricting the statement to the sets F_2^c and $F_1 \cap F_2$. Recall that, as in the previous section, $t_1 = N^2 s_1/L$ and that τ_a is defined before Proposition 6.1.

PROPOSITION 7.1. *Fix $\eta > 0$, $s_1 > 0$ and L, and assume that $Z_{3,a}(0) \leq N$, where N is sufficiently large. Then,*

$$P\big(\tau_a > t_1/4; F_2^c\big) \leq \eta. \tag{7.1}$$

PROPOSITION 7.2. *Fix $\eta > 0$, $s_1 > 0$ and L, and assume that $Z_{3,a}(0) \leq N$, where N is sufficiently large. Then,*

$$P\big(\tau_a > t_1; F_1 \cap F_2\big) \leq \eta. \tag{7.2}$$

Proposition 6.3 follows immediately from Propositions 6.2, 7.1 and 7.2.

The proof of Proposition 7.1 is quick, whereas that of Proposition 7.2 is more involved. The main point in the proof of the former is that the bound $t_1/5L$ is small enough so that, under F_2^c, service at a single class $(3, l, a)$ is fast enough to ensure that $(3, a)$ empties quickly. Here and later on, $c_1, c_2, \ldots$ will denote positive constants whose exact values do not concern us.

PROOF OF PROPOSITION 7.1. Let S_1 denote the sum of the amounts of time spent serving those customers visiting the classes $(3, l, a)$, $l = 1, \ldots, L$, by time $t_1/4$. (This includes the time required for customers there, but not yet served by $t_1/4$.) Under F_2^c, there are at most $t_1/5L$ such customers. The service times at each class $(3, l, a)$ are given by independent exponential random variables with mean $3L/4$. It therefore follows from (5.1) that

$$P(S_1 \geq t_1/5; F_2^c) \leq \exp\{-c_1 N^2\}, \tag{7.3}$$

for appropriate $c_1 > 0$. Moreover, the classes $(3, l, a)$ have priority over the classes $(3, l, b)$. So, when $S_1 < t_1/5 < t_1/4$, the classes $(3, l, a)$ must all be empty at some $t \leq t_1/4$; in particular, $\tau_a \leq t_1/4$. Together with (7.3), this implies that

$$P(\tau_a > t_1/4; F_2^c) \leq \exp\{-c_1 N^2\}. \tag{7.4}$$

The bound in (7.1) follows for large enough N. □

We proceed to the proof of Proposition 7.2. The basic idea is as follows. Under the event F_2, there are a substantial number of customers that visit $(3, a)$ by time $t_1/4$, and hence, typically, a substantial number that visit each class $(3, l, a)$. Under F_1, this includes the class $(3, l_0, a)$, which will be empty at some time in $[t_1/4, t_1/2]$. The customers entering $(3, l_0, a)$ by time $t_1/4$ (or a comparable number of other customers) will then be served at $(3, l_0, b)$ by time $3t_1/4$. Also, these customers will be served at $(2, a)$ by time t_1. This takes time away from the service of customers at $(2, b)$, and hence limits the influx of customers from $(2, b)$ into $(3, a)$. The service at $(3, a)$ will thus be fast enough to ensure that each class $(3, l, a)$ will be empty at some time before t_1, and so $\tau_a \leq t_1$, as desired.

We employ the following lemma as an intermediate step in demonstrating Proposition 7.2. Below, F_3 denotes the event where the time in $[0, t_1]$ spent serving customers at $(2, a)$ is strictly greater than $\gamma t_1/7L^2$.

LEMMA 7.1. *Fix $\eta > 0$, $s_1 > 0$ and $L > 0$. For sufficiently large N,*

$$P(F_1 \cap F_2 \cap F_3^c) \leq \eta. \tag{7.5}$$

PROOF. Let F_4 denote the event where by time $t_1/4$, the total number of customers to visit each of the classes $(3, l, a)$, $l = 1, \dots, L$, is strictly greater than $t_1/6L^2$. Under F_2, the total number of customers to visit $(3, a)$ is strictly greater than $t_1/5L$, and each of these customers chooses a class $(3, l, a)$ randomly. It therefore follows from (5.2) that

$$P(F_2 \cap F_4^c) \leq L \exp\{-c_2 N^2\}, \tag{7.6}$$

for appropriate $c_2 > 0$.

We recall that σ is the first time, after $t_1/4$, at which some class $(3, l, a)$ is empty, and that $(3, l_0, a)$ denotes the class. Under $F_1 \cap F_4$, all of the customers visiting $(3, l_0, a)$, by time σ, are served at $(3, l_0, b)$ before further customers, arriving after σ, are served at $(3, l_0, a)$; there are more than $t_1/6L^2$ of these customers and $\sigma \leq t_1/2$. The mean service time at $(3, l_0, b)$ is γ/L. Since $(\gamma/L)(1/6L^2) < 1/4$, by (5.1), the first $t_1/6L^2$ customers at $(3, l_0, b)$ also typically take at most time $t_1/4$ to serve. Let F_5 denote the event where, by time

$3t_1/4$, the number of customers served in $(3, l_0, b)$ is greater than $t_1/6L^2$. It follows that

$$P(F_1 \cap F_4 \cap F_5^c) \le \exp\{-c_3 N^2\}, \tag{7.7}$$

for appropriate $c_3 > 0$.

The entire time over $(3t_1/4, t_1]$ is available for service of customers at $(2, a)$, of which there are more than $t_1/6L^2$ by time t_1. Since the mean service time at $(2, a)$ is γ, and $\gamma/7L^2 \le 1/4$, another application of (5.1) implies that

$$P(F_3^c \cap F_5) \le \exp\{-c_4 N^2\}, \tag{7.8}$$

for appropriate $c_4 > 0$. Together, (7.6)–(7.8) immediately imply (7.5), for large enough N. □

We now demonstrate Proposition 7.2. This will complete the proof of Proposition 6.3.

PROOF OF PROPOSITION 7.2. On F_3, the amount of time in $[0, t_1]$ available for serving customers in $(2, b)$ is at most $(1 - \gamma/7L^2)t_1$. Denote by F_6 the event that the number of customers in $(2, b)$ served over this time is at most $(4/3)(1 - \gamma/8L^2)t_1$. Since the mean service time at $(2, b)$ is $3/4$, (5.1) implies that

$$P(F_3 \cap F_6^c) \le \exp\{-c_5 N^2\} \tag{7.9}$$

for appropriate $c_5 > 0$.

On F_6, there are at most $(4/3)(1 - \gamma/8L^2)t_1$ customers that enter $(3, a)$ up until time t_1. Let F_7 denote the event that at most $(4/3L)(1 - \gamma/9L^2)t_1$ of these customers enter each of the classes $(3, l, a)$. By (5.2),

$$P(F_6 \cap F_7^c) \le L\exp\{-c_6 N^2\}, \tag{7.10}$$

for appropriate $c_6 > 0$. Including the at most N customers initially at $(3, a)$ and choosing N large enough, at most $(4/3L)(1 - \gamma/10L^2)t_1$ customers visit each class $(3, l, a)$ up until time t_1 on F_7. Let S_2^l, $l = 1, \dots, L$, denote the amount of time spent serving these customers. Since the mean service time of each customer of $(3, a)$ is $3L/4$, (5.1) implies that

$$P(S_2^l \ge (1 - \gamma/11L^2)t_1; F_7) \le \exp\{-c_7 N^2\} \tag{7.11}$$

for $l = 1, \dots, L$, and appropriate $c_7 > 0$.

Together, (7.5) and (7.9)–(7.11) imply that, for a given $\eta > 0$ and a large enough choice of N,

$$P\Big(\max_l S_2^l \ge (1 - \gamma/11L^2)t_1; F_1 \cap F_2\Big) \le \eta. \tag{7.12}$$

But the classes $(3, l, a)$ have priority over the classes $(3, l, b)$. So, if $S_2^l < (1 - \gamma/11L^2)t_1 < t_1$ for a given l, then $(3, l, a)$ must be empty at some t, with $t \le t_1$. In particular, when $\max_l S_2^l < t_1$, then $\tau_a \le t_1$. Together with (7.12), this implies that

$$P(\tau_a > t_1; F_1 \cap F_2) \le \eta, \tag{7.13}$$

which is the bound given in (7.2). □

8. Bounds on the classes $(3, b)$. Here, we obtain upper bounds on the number of customers at each of the classes $(3, l, b)$, $l = 1, \dots, L$. We will show, in Proposition 8.2, that if the numbers of customers at $(3, a)$ and $(3, l, b)$, for a given l, are initially small on the appropriate scale, then the number of customers at $(3, l, b)$ remains comparatively small for a long period of time. These bounds are employed in Section 10 to show that, for the corresponding fluid limits, the amount of mass at $(3, l, b)$ remains 0 for all time for fluid limits starting with no mass there or at $(3, a)$. As before, we require L to be large.

The approach taken here for the classes $(3, l, b)$ differs from the analysis of $(3, l, a)$ in Sections 6 and 7 in several ways. Since the classes $(3, l, a)$ have priority over the classes $(3, l, b)$, it was not necessary to obtain upper bounds on the number of customers at the latter classes for our results in Sections 6 and 7. Here, to analyze the behavior at the classes $(3, l, b)$, we will obviously need to know the behavior at $(3, l, a)$. Fortunately, we can employ Propositions 6.3 and 6.4 for this purpose. On the other hand, the behavior at $(3, b)$ is inherently more elementary than that at $(3, a)$ in the following sense. The rate of service at $(2, b)$ and the combined rate at $(3, a)$ are the same, namely $4/3$, when service is taking place at all of these classes. So, until $(2, a)$ begins service or $(2, b)$ is empty, the evolution of $Z_{3,l,a}(t)$, for a given l, is determined by fluctuations rather than by a net drift. Also, the amount of service required at $(2, a)$ has an important effect on $Z_{3,l,a}(t)$. The evolution of $Z_{3,l,b}(t)$, during individual on and off periods of service, is simpler–linear decrease in the former periods and at most linear increase in the latter. Also, one can show that the on periods will begin, and continue until $(3, l, b)$ is empty, with high probability when the corresponding class $(3, l, a)$ empties, without one needing to analyze much of the network. These differences make the analysis here quicker than in Sections 6 and 7.

Our first result, Proposition 8.1, plays the role of Propositions 6.1 and 6.2. Here, the bounds are additive rather than multiplicative. Also, for application later on, we need estimates at each class $(3, l, b)$, rather than for $(3, b)$. We will employ the following terminology. For $l = 1, \dots, L$, let

$$\tau_{a,l} = \inf\{t\colon Z_{3,l,a}(t) = 0\}.$$

Then, $\tau_a = \max_l \tau_{a,l}$, where τ_a was defined in Section 6. Similarly, we let

$$\tau_{b,l} = \inf\{t\colon Z_{3,l,b}(t) = 0\}.$$

PROPOSITION 8.1. *Fix $\eta > 0$, and suppose that L and N^2/L are sufficiently large. Then, for each $l = 1, \dots, L$,*

$$P\Big(\sup_{t\le N^2} Z_{3,l,b}(t) \ge Z_{3,l,b}(0) + 2N^2/L\Big) \le \eta, \tag{8.1}$$

$$P\Big(\sup_{t\in[\tau_{b,l},\,N^2]} Z_{3,l,b}(t) \ge 2N^2/L\Big) \le \eta \tag{8.2}$$

and

$$P\Big(\sup_{t\in[\tau_{a,l}+N^2/2,\,\tau_{b,l}\wedge N^2]} Z_{3,l,b}(t) \ge Z_{3,l,b}(0) - N^2L\Big) \le \eta. \tag{8.3}$$

PROOF. The bounds in (8.1) and (8.2) are easy to see. Arrivals at $(3, l, b)$, for a given l, are dominated by a rate-$4/3L$ Poisson process. So, (8.1) follows from (5.1), when N^2/L is large. Since $Z_{3,l,b}(\tau_{b,l}) = 0$, the same is true for (8.2).

The bound (8.3) requires more work. Since the reasoning here is similar to that given in the proof of Proposition 6.2, we abbreviate some of the steps. We let $\tau'_{a,l}$ be the first time after $\tau_{a,l}$ at which $(2, a)$ is occupied. When $(3, b, l)$ is occupied at time $\tau_{a,l}$, this occurs with probability at least $1 - 1/L$ before $(3, l, a)$ is again occupied, because of the relative rates of service at $(2, b)$ and $(3, l, b)$. The process $Z_{2,a}(t + \tau'_{a,l})$ dominates a birth–death process with birth rate L/γ and death rate $1/\gamma$, until either $(3, l, b)$ or $(2, a)$ is empty. The probability of the birth–death process ever reaching 0 starting from 1 is $1/L$. So, the probability of service at $(3, l, b)$, over $[\tau_{a,l}, \tau_{b,l}]$, being interrupted by customers at $(3, l, a)$ is at most $2/L$. Since there are no arrivals at $(3, l, b)$ then, it follows that, over $[\tau_{a,l}, \tau_{b,l}]$, $Z_{3,l,b}(\tau_{a,l}) - Z_{3,l,b}(t)$ dominates a rate-L/γ Poisson process, off of the exceptional event of probability $2/L$.

Since $L/\gamma > 6L$, it follows from (5.1), that for large enough N,

$$P\Big(\sup_{t\in[\tau_{a,l}+N^2/2,\,\tau_{b,l}]} Z_{3,l,b}(t) \ge Z_{3,l,b}(\tau_{a,l}) - 3N^2L\Big) \le 3/L. \tag{8.4}$$

Also, off of the exceptional set given in (8.1),

$$Z_{3,l,b}(\tau_{a,l}) < Z_{3,l,b}(0) + 2N^2/L, \tag{8.5}$$

when $\tau_{a,l} \le N^2$. Plugging (8.5) into (8.4), and choosing $L \ge 3/\eta$ implies that (8.3) holds, with a factor of 2 on the right side. Increasing η by a factor of 2 implies (8.3). □

On the set where $\tau_{a,l} \le N^2/2$, the inequalities (8.2) and (8.3) together show that

$$Z_{3,l,b}(N^2) < \big(Z_{3,l,b}(0) - N^2L\big) \vee (2N^2/L)$$

typically holds. For $s_1/L \le 1/2$, one has $t_1 \le N^2/2$. So, by employing Proposition 6.3, one can remove this restriction on $\tau_{a,l}$ in the following result. Note that application of Proposition 6.3 introduces the condition $Z_{3,a}(0) \le N$ here.

COROLLARY 8.1. *For a given $\eta > 0$, let L be sufficiently large. Assume that $Z_{3,a}(0) \le N$, where N is sufficiently large. Then, for each $l = 1, \dots, L$,*

$$P\big(Z_{3,l,b}(N^2) \ge \big(Z_{3,l,b}(0) - N^2L\big) \vee (2N^2/L)\big) \le \eta. \tag{8.6}$$

We employ the bounds (8.1) and (8.6), in conjunction with Propositions 5.1 and 6.4, to Proposition 8.2. The proposition states that if the numbers of customers at $(3, a)$ and $(3, l, b)$, for a given l, are initially at most given powers of M, then this will typically continue to hold at $(3, l, b)$ for a long period of time relative to M.

PROPOSITION 8.2. *For given r_0 and r_1, with $0 < 2r_0 < r_1$, let L be sufficiently large. Assume that $Z_{3,a}(0) \le M^{r_0}$ and $Z_{3,l,b}(0) \le M^{r_1}/2$, for given l, where M is sufficiently large. Then,*

$$P\big(Z_{3,l,b}(t) \ge M^{r_1} \textit{ for some } t \le M^2\big) \le 2/M. \tag{8.7}$$

For our purposes, $0 < 2r_0 < r_1 < 1$ gives the relevant ranges of r_0 and r_1. When analyzing the fluid limits of $Z(t)$ in Section 10, we will scale time and $Z(t)$ by M. The above bound on r_1 is therefore small enough to ensure that the number of customers at $(3, l, b)$ scales to 0. We will only need the proposition when $Z_{3,l,b}(0) = 0$. One can, on the other hand, check that when $L = 1$, (8.7) can fail even when $Z_3(0) = 0$. [If service at $(2, b)$ continues through time t, the return times of $Z_{3,a,l}(s)$ to 0, for a given l, and thus the variable $Z_{3,b,l}(s)$, will repeatedly be of order t on $[0, t]$.]

PROOF OF PROPOSITION 8.2. Choose $r' \in (r_0, r_1/2)$. By (6.14) of Proposition 6.4 and the assumption $Z_{3,a}(0) \le M^{r_0}$,

$$P\big(Z_{3,a}(t) \ge M^{r'} \text{ for some } t \le M^2\big) \le 1/M. \tag{8.8}$$

Define the process $\widetilde{Z}(t)$ by setting $\widetilde{Z}(t) = Z(t)$ until $Z_{3,a}(t) \ge M^{r'}$, and after which setting $\widetilde{Z}(t) = \Delta$, where $\{\Delta\}$ is an extension of the state space. For $\widetilde{Z}(t) = \Delta$, also set $\widetilde{Z}_{3,l,b}(t) = 0$. By (8.8), to show (8.7), it suffices to show instead

$$P\big(\widetilde{Z}_{3,l,b}(t) \ge M^{r_1} \text{ for some } t \le M^2\big) \le 1/M, \tag{8.9}$$

for $\widetilde{Z}_{3,l,b}(0) \le M^{r_1}/2$ and large enough M.

Define $\|\cdot\|$ so that $\|\widetilde{Z}(t)\| = \widetilde{Z}_{3,l,b}(t)/M^{2r'}$. It is not difficult to verify that the conditions of Proposition 5.1 are satisfied for $\widetilde{Z}(t)$ and $\|\cdot\|$. Since $Z(t)$ is a Markov process with jump rates at most $\Gamma = 6L^2/\gamma$ and upward jumps at most $J = 1$, the same is true for $\widetilde{Z}(t)$. For given $\widetilde{Z}(0)$, consider the probability that

$$\|\widetilde{Z}(t)\| - \|\widetilde{Z}(0)\| = \big(\widetilde{Z}_{3,l,b}(t) - \widetilde{Z}_{3,l,b}(0)\big)/M^{2r'}$$

first exits the interval $(-1, 1)$ on the right. Setting $N = M^{r'}$ in (8.1) and (8.6), it follows that, for sufficiently large L and M, this probability can be chosen as close to 0 as desired, when $\|\widetilde{Z}(0)\| \geq 1$. So, for given $\eta > 0$, and large L and M, (5.10) is satisfied by $\widetilde{Z}(t)$ and $\|\cdot\|$, with $R_1 = R_2 = \nu = 1$.

It follows from (5.11) of Proposition 5.1, that for $r = 1$ and large enough L,

$$P\big(\|\widetilde{Z}(t)\| \geq M + N \text{ for some } t \leq e^M\big) \leq e^{-M} \tag{8.10}$$

holds for $\|\widetilde{Z}(0)\| \leq N$, with $N \geq 1$, and large enough M. Substitution of $M^{r_1 - 2r'}/2$ for both M and N in (8.10) shows that (8.9) holds for $\widetilde{Z}_{3,l,b}(0) \leq M^{r_1}/2$ and large enough M. This, in turn, implies (8.7). □

9. Restarting the process. In Section 10, we will apply Propositions 6.4 and 8.2 to the stations $(3, l)$, $l = 1, \ldots, L$, of the queueing network in (2.1), (2.2). These results assume initial data at $(3, l)$ which are small on an appropriate scale; they then state that the number of customers there will typically remain comparatively small for long times. Proposition 6.4 does this for the classes $(3, l, a)$, and Proposition 8.2 does this for $(3, l, b)$. Since these assumptions need not hold for $Z(0)$, we need to show that we can find appropriate stopping times κ_a and $\kappa_{b,l}$, $l = 1, \ldots, L$, where they typically do hold. This is the purpose of the current section.

The main results here are Propositions 9.3 and 9.4, with the former treating $(3, l, a)$, and the latter treating $(3, l, b)$. Both κ_a and $\kappa_{b,l}$ will be at most fixed multiples of $|Z(0)|$. In Proposition 9.3, we will show that $Z_{3,l,a}(\kappa_a)$ is small simultaneously for all l, whereas, in Proposition 9.4, the argument $\kappa_{b,l}$ in $Z_{3,l,b}(\kappa_{b,l})$ is allowed to vary with l. On account of this, Proposition 9.4 will be easy to show, whereas the demonstration of Proposition 9.3 will occupy most of the section.

We proceed to analyze $Z_{3,l,a}(t)$ by reinterpreting the arrivals and departures at $(3, l, a)$, $l = 1, \ldots, L$. Arrivals at each $(3, l, a)$ are due to departures at $(2, b)$. These can only occur when $Z_{2,b}(t) > 0$ but $Z_{2,a}(t) = 0$, during which periods they are given by independent Poisson processes, with intensity $4/3L$ for each l. We may therefore construct L independent Poisson processes, each with intensity $4/3L$, corresponding to all *potential arrivals* at each l. *Actual arrivals* at $(3, l, a)$ occur when $Z_{2,b}(t) > 0$ and $Z_{2,a}(t) = 0$; *ghost arrivals* occur when at least one of these two conditions fails. So, $Z_{2,b}(t)$ decreases by 1 and $Z_{3,l,a}(t)$ increases by 1 when an actual arrival occurs, and otherwise remain the same. Similarly, *potential departures* at a given $(3, l, a)$ occur according to independent Poisson processes, each with intensity $4/3L$. These are given by *actual departures* or *ghost departures*, depending on whether or not $Z_{3,l,a}(t) > 0$. When an actual departure occurs, $Z_{3,l,a}(t)$ decreases by 1 and $Z_{3,l,b}(t)$ increases by 1.

In order to analyze $Z_{3,l,a}(t)$, for a given l, we first compare it with the process $X_l^1(t)$, which is identical to $Z_{3,l,a}(t)$, except that ghost departures at $(3, l, a)$ are included. That is, $X_l^1(t)$ decreases by 1 whenever a potential departure occurs. The process $Z_{3,l,a}(t)$ is obtained from $X_l^1(t)$ by "reflecting"

$X_l^1(t)$ at 0. So, one can write $Z_{3,l,a}(t)$ in terms of $X_l^1(t)$, by the standard pathwise formula

$$Z_{3,l,a}(t) = X_l^1(t) - \Big(\inf_{s \le t} X_l^1(x)\Big) \wedge 0. \tag{9.1}$$

In order to analyze $X_l^1(t)$, we compare it with the process $X_l(t)$, which is identical to $X_l^1(t)$, except that ghost arrivals at $(3, l, a)$ are included. That is, $X_l(t)$ increases by 1 whenever a potential arrival occurs. Since $X_l(t)$ is a continuous time symmetric nearest neighbor random walk, with upward and downward jumps each occurring at rate $4/3L$, it is easily analyzed, as in Proposition 9.1. We can write

$$X_l(t) = X_l^1(t) + X_l^2(t), \tag{9.2}$$

where $X_l^2(t)$ is the process of ghost arrivals at $(3, l, a)$. The station 2 is strictly subcritical, and so it is not difficult to show, as in Lemma 9.1, that $Z_2(t) = 0$ at least a fixed portion of the time. This provides lower bounds on $X_l^2(t)$, as in Lemma 9.2. Together with the bounds on $X_l(t)$ and (9.2), this will provide upper bounds on $X_l^1(t)$ at select random times, as in Proposition 9.2, and hence, using (9.1), on $Z_{3,l,a}(t)$ at the same times. Since $X_l^2(t)$, $l = 1, \dots, L$, will tend to grow in unison when $Z_{2,b}(t) = 0$ or $Z_{2,a}(t) > 0$, this procedure, in fact, provides bounds which hold simultaneously on $Z_{3,l,a}(t)$, for all l, at select random times, as in Proposition 9.3.

In contrast to the above procedure for bounding $Z_{3,l,a}(t)$ simultaneously for all l, the procedure we employ for bounding $Z_{3,l,b}(t)$, for a given l, requires almost no work. This follows from the strict subcriticality of station $(3, l)$, which ensures that it will be empty at some random time which is typically at most a fixed multiple of $|Z(0)|$.

We now begin our analysis of $Z_{3,l,a}(t)$. Let

$$\mathscr{E}_j(t) = \big|\{s\colon Z_j(s) = 0, s \le t\}\big|$$

be the amount of time up until t during which the station j is empty. The following lemma is a consequence of the strict subcriticality of the individual stations in the queueing network. Here, we set $|Z(0)| = z$.

LEMMA 9.1. *For large enough t satisfying $t \ge 32z$ and appropriate $c_8 > 0$,*

$$P\big(\mathscr{E}_j(t) \le t/16\big) \le \exp\{-c_8 t\} \tag{9.3}$$

when $j = 1, 2$. When $j = (3, l)$, $l = 1, \dots, L$, (9.3) *holds for large enough t satisfying $t \ge 32Lz$.*

PROOF. The argument is straightforward, since each class is visited at most once by a given customer. The sum of the mean service times at each station j, $j = 1, 2$, is less than 7/8 by (2.2). Moreover, by (5.1), the total number of distinct customers in the network up until time t will be, for fixed $\varepsilon > 0$, at most $(1 + \varepsilon)t + z$ off of a set of exponentially small probability. Setting $\varepsilon = 1/32$ and substituting for z gives the upper bound $17t/16$. Since

$(17/16)(7/8) < 15/16$, (9.3), for $j = 1, 2$, follows from another application of (5.1).

The reasoning for $j = (3, l)$ is the same, except that one requires both (5.1) and (5.2) so that, off of a set of exponentially small probability, at most $((1+\varepsilon)/L)t + z$ customers visit $(3, l)$, for a given l. The sum of the mean service times at $(3, l, a)$ and $(3, l, b)$ is less than $7L/8$ by (2.2). The reasoning then continues as before, since the factors of L cancel. □

Denote the first time at which the station j is empty by τ_j. The following is an immediate consequence of Lemma 9.1.

COROLLARY 9.1. *For large enough t satisfying $t \geq 32z$ when $j = 1, 2$, and $t \geq 32Lz$ when $j = (3, l)$, $l = 1, \dots, L$,*

$$P(\tau_j > t) \leq \exp\{-c_8 t\} \tag{9.4}$$

for appropriate $c_8 > 0$.

We wish to consider the behavior of $Z_2(s)$ on $(t, 35t]$, for large t with $t \geq 32Lz$. (Later on, we will apply Corollary 9.1 at t with $j = (3, l)$, for each l.) We note that, off of a set of exponentially small probability in t, $|Z(t)| \leq 33t/32 + z \leq 17t/16$. So, one can apply Lemma 9.1 to the process restarted at time t, with $j = 2$. By (9.3), it follows that on $(t, 35t]$, the proportion of time that $Z_2(s) = 0$ is at least $1/16$ off of a set of exponentially small probability. We decompose this interval into $n_t = \lfloor 34t^{2/7} \rfloor$ subintervals $I_i = (t_{i-1}, t_i]$, where $t_i = t + it^{5/7}$ and $i = 1, \dots, n_t$, leaving off the last piece of length less than $t^{5/7}$. [The choice of the power 5/7 is somewhat arbitrary; any choice in (1/2, 1) will suffice.] It is then easy to see that, off of the exceptional random set, for at least one such interval, the proportion of time that $Z_2(s) = 0$ is at least $1/20$. Denote the first such interval by I_{i_0}, when it exists, and the presence of such an interval by $i_0 < \infty$. For later use, we set

$$\kappa_a = \begin{cases} t_{i_0}, & \text{for } i_0 < \infty, \\ 35t, & \text{for } i_0 = \infty, \end{cases}$$

where κ_a is the first of the two stopping times mentioned at the beginning of the section. Note that $t \leq \kappa_a \leq 35t$ always holds. We have shown the following.

COROLLARY 9.2. *For large enough t satisfying $t \geq 32Lz$ and appropriate $c_9 > 0$,*

$$P(i_0 = \infty) \leq \exp\{-c_9 t\}. \tag{9.5}$$

One of the processes employed to analyze $Z_{3,l,a}(t)$, $l = 1, \dots, L$, was $X_l(t)$, the rate-$8/3L$ continuous time symmetric nearest neighbor random walk introduced below (9.1). The reversed process $\widetilde{X}_l(s) = X_l(T) - X_l(T-s)$, with $s \in [0, T]$ and T fixed, is also a rate-$8/3L$ continuous time symmetric nearest neighbor random walk. Applying (5.3) to $\widetilde{X}_l(s)$, with $\beta = 4/7$ and $\beta = 3/5$, respectively, we obtain the following bounds on the fluctuations of $X_l(s)$, measured backwards from T.

PROPOSITION 9.1. *Let $T \le 35t$, with t sufficiently large. For each l and appropriate $c_{10} > 0$,*

$$\begin{aligned} &P\left(X_l(T) - \inf\{X_l(s)\colon s \le T\} \ge t^{4/7}\right) \le \exp\{-c_{10}t^{1/7}\}, \\ &P\left(X_l(T) - \inf\{X_l(s)\colon T - t^{5/7} \le s \le T\} \ge t^{3/7}\right) \le \exp\{-c_{10}t^{1/7}\}. \end{aligned} \tag{9.6}$$

We will need to control the fluctuations of $X_l(s)$ corresponding to (9.6), but with T replaced by the random time κ_a defined above. On $i_0 < \infty$, κ_a can take only n_t values. Applying Proposition 9.1 and adding up the probabilities of a large fluctuation for each $i = 1, \dots, n_t$, one obtains analogous estimates at time κ_a.

COROLLARY 9.3. *Let t be sufficiently large. For each l and appropriate $c_{10} > 0$,*

$$\begin{aligned} &P\left(X_l(\kappa_a) - \inf\{X_l(s)\colon s \le \kappa_a\} \ge t^{4/7}; i_0 < \infty\right) \\ &\qquad \le n_t \exp\{-c_{10}t^{1/7}\}, \\ &P\left(X_l(\kappa_a) - \inf\{X_l(s)\colon \kappa_a - t^{5/7} \le s \le \kappa_a\} \ge t^{3/7}; i_0 < \infty\right) \\ &\qquad \le n_t \exp\{-c_{10}t^{1/7}\}. \end{aligned} \tag{9.7}$$

Denote by V_i the amount of time that station 2 is empty over the intervals I_i, $i = 1, \dots, n_t$, specified above. Also, denote by $X^2_{l,i}$ the number of ghost arrivals at $(3, l, a)$ over I_i. Potential arrivals at $(3, l, a)$ occur independently of the state of the queueing network, according to a Poisson process with rate $4/3L$. Ghost arrivals occur there at the same rate when station 2 is empty. [They also occur when $(2, a)$ is occupied, although we will not use this here.] So, on the set $\{V_i \ge t^{5/7}/20\}$, $X^2_{l,i}$ dominates a Poisson random variable with mean $t^{5/7}/15L$. Using the large deviation estimate (5.1), one obtains the following bounds on $X^2_{l,i}$.

LEMMA 9.2. *For each $i = 1, \dots, n_t$ and l, and large enough t,*

$$P\left(X^2_{l,i} \le t^{5/7}/20L; V_i \ge t^{5/7}/20\right) \le \exp\{-c_{11}t^{5/7}\} \tag{9.8}$$

for appropriate $c_{11} > 0$.

Applying these bounds to $i = i_0$ implies the following.

COROLLARY 9.4. *For each l and large enough t,*

$$P\left(X^2_{l,i_0} \le t^{5/7}/20L; i_0 < \infty\right) \le n_t \exp\{-c_{11}t^{5/7}\} \tag{9.9}$$

for appropriate $c_{11} > 0$.

Recall that the process $X_l(s)$ of potential arrivals at $(3, l, a)$ is the sum of the processes $X_l^1(s)$ and $X_l^2(s)$ of actual and ghost arrivals. When $i_0 < \infty$ and $X_{l,i_0}^2 > t^{5/7}/20L$, it follows that for $s \le t_{i_0-1}$,

$$\begin{aligned} X_l^1(\kappa_a) - X_l^1(s) &= \big(X_l(\kappa_a) - X_l(s)\big) - \big(X_l^2(\kappa_a) - X_l^2(s)\big) \\ &\le \big(X_l(\kappa_a) - X_l(s)\big) - X_{l,i_0}^2 \\ &< X_l(\kappa_a) - X_l(s) - t^{5/7}/20L. \end{aligned} \tag{9.10}$$

This allows us to control $X_l^1(\kappa_a) - X_l^1(s)$ on the complement of the event in (9.9).

Proposition 9.2 gives bounds on the fluctuations of $X_l^1(s)$, measured backwards from κ_a. For $s \in (\kappa_a - t^{5/7}, \kappa_a]$, the result is a direct application of the second part of (9.7) and the monotonicity of $X_l^2(s)$. For $s \in [0, \kappa_a - t^{5/7}]$, one instead needs the first part of (9.7), together with (9.9) and (9.10).

PROPOSITION 9.2. *For each l and large enough t,*

$$P\Big(X_l^1(\kappa_a) - \inf_{s \le \kappa_a} X_l^1(s) \ge t^{3/7}; i_0 < \infty\Big) \le 3n_t \exp\{-c_{12}t^{1/7}\} \tag{9.11}$$

for appropriate $c_{12} > 0$.

Employing Proposition 9.2 together with the previous estimates, it is now straightforward to show Proposition 9.3. By Corollary 9.2, for large t satisfying $t \ge 32Lz$, $i_0 < \infty$ occurs off of a set of (exponentially) small probability. Also, by Corollary 9.1, $\tau_{3,l} \le t \le \kappa_a$ off of a set of small probability. By (9.1) and Proposition 9.2, on $i_0 < \infty$ and $\tau_{3,l} \le \kappa_a$, one has $Z_{3,l,a}(\kappa_a) < t^{3/7}$ off of a set of small probability. Together, these results give the desired behavior of $Z_{3,l,a}(\kappa_a)$.

PROPOSITION 9.3. *For each l and large enough t, with $t \ge 32Lz$,*

$$P\big(Z_{3,l,a}(\kappa_a) \ge t^{3/7}\big) \le \exp\{-c_{13}t^{1/7}\} \tag{9.12}$$

for appropriate $c_{13} > 0$.

Recall that κ_a satisfies $\kappa_a \le 35t$, and does not depend on l.

We also want to show the analog of (9.12) for the classes $(3, l, b)$, at appropriate times $\kappa_{b,l} \in [35t, dt]$, for appropriate d. The reasoning is, in this case, much simpler. Again, assume that $t \ge 32Lz$. Customers enter the network at rate 1, and so by (5.1), off of a set of exponentially small probability, $|Z(35t)| \le 36t + z \le 37t$. Restarting the process at this time, we let $\tau'_{3,l}$ denote the first time at which the station $(3, l)$ of the restarted process is empty. By Corollary 9.1, off of a set of exponentially small probability, $\tau'_{3,l} \le 32 \cdot 37Lt$. Set $d = 35 + (32 \cdot 37)L$, and let $\kappa_{b,l}$ be the stopping time

$$\kappa_{b,l} = \big(35t + \tau'_{3,l}\big) \wedge (dt).$$

We thus obtain the following result. Note that $\kappa_a \le \kappa_{b,l} \le dt$.

PROPOSITION 9.4. *For each l and large enough t, with $t \geq 32Lz$,*

$$P\big(Z_{3,l,b}(\kappa_{b,l}) > 0\big) \leq \exp\{-c_{14}t\} \tag{9.13}$$

for appropriate $c_{14} > 0$.

10. Proof of Theorem 3.

In this section, we complete the proof of Theorem 1, that the Markov process $Z(t)$ for the network (2.1), (2.2) is positive recurrent. We recall from Section 4 that it suffices to demonstrate Theorem 3. For this, in turn, it suffices to verify that the fluid limit model corresponding to $Z(t)$ is asymptotically stable.

As summarized in Section 4, there are several main steps in showing the fluid limit model is asymptotically stable. Most of the work has already been done in Sections 5–9. In particular, using Propositions 6.4, 8.2, 9.3 and 9.4, it will follow that any fluid limit $(\bar{T}(t), \bar{Z}(t))$ satisfies $\bar{Z}_3(t) = 0$ for large enough t. This is shown in Proposition 10.1. One can therefore, in effect, omit station 3 when analyzing the behavior of the fluid limits. For the resulting network, it suffices to analyze the behavior of the corresponding fluid model equations. If one omits station 3, the network is strictly subcritical with route

$$\to (1,b) \to (2,b) \to (2,a) \to (1,a) \to \tag{10.1}$$

and mean service times

$$m_{1,a} = m_{2,b} = \tfrac{3}{4}, \qquad m_{1,b} = m_{2,a} = \gamma, \tag{10.2}$$

with $\gamma \in (0, 1/8)$. This reduced network is last-buffer-first-served; by Dai and Weiss (1996), its fluid model is stable. This enables us to show, in Proposition 10.3, that our fluid limit model is asymptotically stable.

We now tie together our results from Sections 5–9 to show $\bar{Z}_3(t) = 0$ for large t.

PROPOSITION 10.1. *For sufficiently large L and appropriate events H^z, with $P(H^z) \to 1$ as $|z| \to \infty$, all fluid limits $(\bar{T}(t), \bar{Z}(t))$ on H^z, of $Z(t)$, satisfy*

$$\bar{Z}_3(t) = 0 \quad \textit{for } t \geq t_0 \tag{10.3}$$

and appropriate t_0. Moreover,

$$\big|\bar{Z}(t_0)\big| \leq 2t_0 + 1. \tag{10.4}$$

PROOF. We first recall the stopping times κ_a and $\kappa_{b,l}$, $l = 1, \ldots, L$, employed in Propositions 9.3 and 9.4. Choosing $t = 32L|z|$ in both places, one has

$$\kappa_a \leq \kappa_{b,l} \leq t_0|z| \tag{10.5}$$

for $t_0 = 32dL$, where d is given at the end of Section 9. By Proposition 9.3, for large $|z|$,

$$P\big(Z_{3,a}(\kappa_a) \geq L(32L|z|)^{3/7}\big) \leq L\exp\big\{-c_{13}(32L|z|)^{1/7}\big\}, \tag{10.6}$$

and by Proposition 9.4, for large z and given l,

$$P\big(Z_{3,l,b}(\kappa_{b,l}) > 0\big) \le \exp\{-c_{14}|z|\}, \tag{10.7}$$

where $c_{13} > 0$ and $c_{14} > 0$.

Restart the process $Z(t)$ at κ_a. By (10.6), Proposition 6.4 and the strong Markov property, for sufficiently large $|z|$,

$$P\big(Z_{3,a}(t) \ge |z|^{4/9} \text{ for some } t \in [\kappa_a, |z|^2]\big) \le 2/|z|. \tag{10.8}$$

[In Proposition 6.4, we are setting $M = |z|$, $r_0 \in (3/7, 4/9)$ and $r_1 = 4/9$.] In particular, by (10.5), this includes $t = \kappa_{b,l}$, for $|z|$ large enough and given l.

Restart $Z(t)$ again, this time at $\kappa_{b,l}$. By (10.7), (10.8), Proposition 8.2 and the strong Markov property,

$$P\big(Z_{3,l,b}(t) \ge |z|^{9/10} \text{ for some } t \in \big[\kappa_{b,l}, |z|^2\big]\big) \le 5/|z|. \tag{10.9}$$

(In Proposition 8.2, set $M = |z|$, $r_0 = 4/9$ and $r_1 = 9/10$.) By (10.5), (10.8) and (10.9), one obtains that

$$P\big(Z_3(t) \ge (L+1)|z|^{9/10} \text{ for some } t \in \big[t_0|z|, |z|^2\big]\big) \to 0 \tag{10.10}$$

as $|z| \to \infty$. Let H_1^z denote the complement of the exceptional set in (10.10). Clearly, each fluid limit $(\bar{T}(t), \bar{Z}(t))$ on H_1^z satisfies (10.3), and $P(H_1^z) \to 1$ as $|z| \to \infty$.

We still need to bound $|\bar{Z}(t_0)|$. Let $E(t)$ denote the number of external arrivals into the network [at $(1, b)$] by time t. This does not depend on the initial state z. By the weak law of large numbers, $E(t)/t \to 1$, in probability, as $t \to \infty$. Since $Z(t)$ only increases through external arrivals, it follows that

$$P\left(\frac{1}{|z|}|Z(t_0|z|)| > 2t_0 + 1\right) \to 0 \quad \text{in probability}, \tag{10.11}$$

as $|z| \to \infty$. Consequently, (10.4) holds for each fluid limit on H_2^z, where H_2^z is the complement of the exceptional set in (10.11). Setting $H^z = H_1^z \cap H_2^z$, one has $P(H^z) \to 1$ as $|z| \to \infty$, which completes the proof of the proposition. □

In the following lemma, we show that fluid model solutions $(\bar{T}(t), \bar{Z}(t))$ for the original network (2.1), (2.2) which satisfy $\bar{Z}_3(t) \equiv 0$ are also fluid model solutions for the reduced network (10.1), (10.2). To simplify notation, we set $\bar{D}_k(t) = m_k^{-1}\bar{T}_k(t)$ for all classes k, and let $\bar{D}_{3,a}(t)$ and $\bar{D}_{3,b}(t)$ denote the corresponding sums over l. [$\bar{D}_k(t)$ measures the "departures" at k.]

LEMMA 10.1. *Let* $(\bar{T}(t), \bar{Z}(t))$ *be a solution of the fluid model equations* (2.10), (2.11) *for the network* (2.1), (2.2), *with* $\bar{Z}_3(t) = 0$ *for all* t. *Then, the restriction of* $(\bar{T}(t), \bar{Z}(t))$ *to the stations* 1 *and* 2 *is a solution of* (2.10), (2.11) *for the network* (10.1), (10.2).

PROOF. The equations (2.10) and (2.11) for the network (10.1), (10.2) follow automatically from their analogs for the network (2.1), (2.2), except for

$$\begin{aligned}\bar{Z}_{2,a}(t) &= \bar{Z}_{2,a}(0) + m_{2,b}^{-1}\bar{T}_{2,b}(t) - m_{2,a}^{-1}\bar{T}_{2,a}(t)\\ &= \bar{Z}_{2,a}(0) + \bar{D}_{2,b}(t) - \bar{D}_{2,a}(t),\end{aligned} \tag{10.12}$$

which needs to be shown. By (2.10) of the original network,

$$\bar{Z}_{3,a}(t) = \bar{Z}_{3,a}(0) + \bar{D}_{2,b}(t) - \bar{D}_{3,a}(t),$$

which implies that $\bar{D}_{2,b}(t) = \bar{D}_{3,a}(t)$ for all t, since $\bar{Z}_{3,a}(t) = 0$. Similarly, $\bar{D}_{3,a}(t) = \bar{D}_{3,b}(t)$. So, $\bar{D}_{2,b}(t) = \bar{D}_{3,b}(t)$ for all t. Also, by (2.10),

$$\bar{Z}_{2,a}(t) = \bar{Z}_{2,a}(0) + \bar{D}_{3,b}(t) - \bar{D}_{2,a}(t). \tag{10.13}$$

Together with the previous equality, (10.13) implies (10.12). □

The discipline of the network given in (10.1), (10.2) is last-buffer-first-served (LBFS). Since the traffic intensity at each station is less than 1, the network is strictly subcritical. We can therefore employ the following result, Theorem 4.4 from Dai and Weiss (1996).

PROPOSITION 10.2. *The fluid model corresponding to any strictly subcritical network with the LBFS discipline is stable.*

Together, Propositions 10.1 and 10.2, and Lemma 10.1 imply that the fluid limit model corresponding to the process $Z(t)$ is asymptotically stable. This demonstrates Theorem 3, and hence completes the proof that $Z(t)$ is positive recurrent.

PROPOSITION 10.3. *For sufficiently large L, the fluid limit model corresponding to the process $Z(t)$ is asymptotically stable.*

PROOF. Choose the events H^z as in Proposition 10.1. Then, $P(H^z) \to 1$ as $|z| \to \infty$, and all fluid limits $(\bar{T}(t), \bar{Z}(t))$ on H^z satisfy $\bar{Z}_3(t) = 0$ for all $t \geq t_0$, with

$$\left|\bar{Z}(t_0)\right| \leq 2t_0 + 1. \tag{10.14}$$

Set $\widetilde{Z}(t) = \bar{Z}(t + t_0)$. Then, $\widetilde{Z}(t)$ is a solution of the fluid model equations (2.10), (2.11), with $\widetilde{Z}_3(t) = 0$ for all t. By Lemma 10.1, the restriction of $\widetilde{Z}(t)$ to the stations 1 and 2 is a solution of (2.10), (2.11) for the reduced network (10.1), (10.2). This network is strictly subcritical, with $\rho_1 = \rho_2 = 3/4 + \gamma < 1$, and its discipline is LBFS. So, by Proposition 10.2, the restriction of $\widetilde{Z}(t)$ is stable; that is, $\widetilde{Z}_1(t) = 0$ and $\widetilde{Z}_2(t) = 0$ for $t \geq t_1|\widetilde{Z}(0)|$, and appropriate t_1. Consequently,

$$\widetilde{Z}(t) = 0 \quad \text{for } t \geq t_1\left|\widetilde{Z}(0)\right|. \tag{10.15}$$

Converting back to $\bar{Z}(t)$ and employing (10.14), one obtains from (10.15) that

$$\bar{Z}(t) = 0 \quad \text{for } t \geq \delta, \tag{10.16}$$

for $\delta = t_0 + t_1(2t_0 + 1)$. Since this holds for all fluid limits on H^z, the fluid limit model corresponding to $Z(t)$ is asymptotically stable, as desired. □

Acknowledgment. The author thanks V. Dumas for illuminating discussions on the applications of fluid limits.

REFERENCES

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

BRAMSON, M. (1994). Instability of FIFO queueing networks. *Ann. Appl. Probab.* **4** 414–431.

BRAMSON, M. (1998). Stability of two families of queueing networks and a discussion of fluid limits. *Queueing Systems Theory Appl.* **28** 7–31.

DAI, J. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49–77.

DAI, J. (1996). A fluid-limit model criterion for instability of multiclass queueing networks. *Ann. Appl. Probab.* **6** 751–757.

DAI, J. and MEYN, S. (1995). Stability and convergence of moments of multiclass queueing networks via fluid models. *IEEE Trans. Automat. Control* **40** 1889–1904.

DAI, J. and WEISS, G. (1996). Stability and instability of fluid models for re-entrant lines. *Math. Oper. Res.* **21** 115–134.

ETHIER, S. and KURTZ, T. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.

LU, S. H. and KUMAR, P. R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Automat. Control* **36** 1406–1416.

MEYN, S. (1995). Transience of multiclass queueing networks via fluid limit models. *Ann. Appl. Probab.* **5** 946–957.

RYBKO, S. and STOLYAR, A. (1992). Ergodicity of stochastic processes that describe the functioning of open queueing networks. *Problems Inform. Transmission* **28** 3–26 (in Russian).

SEIDMAN, T. I. (1994). "First come, first served" can be unstable! *IEEE Trans. Automat. Control* **39** 2166–2171.

STOLYAR, A. (1994). On the stability of multiclass queueing networks. In *Proceedings of the Second International Conference on Telecommunication Systems—Modeling and Analysis, Nashville, TN* 1020–1028.

SCHOOL OF MATHEMATICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455
E-MAIL: bramson@math.umn.edu