from Queueing and Related Models
U.N. Bhat and I.V. Basawa (editors)
Oxford Science Publications
Clarendon Press, Oxford 1992, pp. 112-12

7
A tandem fluid network with Lévy input

Offer Kella and Ward Whitt

ABSTRACT

We introduce an open network fluid model with stochastic input and deterministic linear internal flows. In particular, we consider several buffers with unlimited capacity in series. The input to the first buffer is a non-decreasing stochastic process with stationary and independent increments. The content flows forward from buffer to buffer through connecting pipes at constant deterministic rates. We obtain simple expressions for the mean content of each buffer and each pipe by exploiting a connection to the classical single-node storage model with non-decreasing Lévy input and constant release rate. We obtain the marginal distributions describing the content of each buffer by exploiting a connection to a linear fluid model with random disruptions. We apply martingale theory to derive the joint distribution of the content of the first two buffers, which is not of product form. Finally, we show that the fluid network can be regarded as the limit of a sequence of conventional queueing networks.

7.1 INTRODUCTION

It is hard to fathom the passage of time. It almost seems like yesterday, but it was twenty-six years ago at Cornell that the second author took Professor Prabhu's course in queueing theory, based on his then recently completed book (Prabhu, 1965). Surprisingly, Professor Prabhu seems much the same today as he did then, a dedicated scholar with a quiet dignity. We hope that Professor Prabhu takes satisfaction from the fact that quite a few of his former students have continued to work in the same field. We respectfully dedicate this chapter to him.

Two of Professor Prabhu's favourite topics over the years have been stochastic storage models and Lévy processes. Indeed, these topics have a prominent place in his two queueing books (Prabhu, 1965, 1980). It thus seems appropriate that these two topics should be the focus of the present paper. Indeed, the purpose of the present paper is to consider a network generalization of the classical storage model with nondecreasing Lévy input and constant release rate; see Chapter 7 of Prabhu (1965) and Chapter 3 of Prabhu (1980).

The present paper can also be viewed in relation to the closely related growing literature on deterministic fluid models; e.g., see Newell (1982), Anick *et al.* (1982), Mitra (1988), Chen and Mandelbaum (1991) and Chen and Yao (1992). Since deterministic fluid flow is usually easier to analyze than a corresponding stochastic process, the deterministic fluid models often provide ways to effectively analyze complicated congestion systems. As in Anick *et al.* (1982), Chen and Yao (1992), and Mitra (1988), a good way to represent stochastic behaviour occurring in different time scales is to combine deterministic fluid flow with stochastic features. The deterministic fluid flow then represents features whose random fluctuations are in a shorter time scale than the stochastic features of the model. The law of large numbers helps justify approximating the features whose random fluctuations are in a shorter time scale by deterministic fluid flow in the larger time scale.

In this paper we introduce and investigate an open network fluid model with stochastic input that may have useful applications and is relatively easy to analyse. Since only the input is stochastic, this model is a natural candidate when the random fluctuations within the network occur in a shorter time scale than the random fluctuations in the input.

In particular, motivated by the evolution of high-speed communication networks that will carry diverse traffic, including very long messages such as long file transfers and video, as well as short signalling messages, we suggest considering an open network model with non-decreasing Lévy process input at some of the nodes and deterministic fluid flow within the network. Such a network storage model can be defined in terms of primitive data consisting of a vector Lévy input process, a routing matrix, and a vector of flow rates by applying the multi-dimensional reflection map in Harrison and Reiman (1981) and Chen and Mandelbaum (1991) (which is shown to be Lipschitz continuous on the function space D with the Skorohod (1956) J_1 and M_1 topologies by Chen and Whitt, 1991). We intend to study this general stochastic network elsewhere. In this paper we only consider the special case of a tandem fluid network (i.e. n nodes in series). In particular, here all nodes have unlimited capacity, there is exogenous Lévy process input only at the first node and there is linear deterministic flow forward through successive nodes. This model might represent one communication path through a communication network.

From some perspectives, this tandem fluid model is remarkably easy to analyse, because for each k the total content at the first k nodes (after adjusting for propagation delay) behaves exactly like the classical single-node storage model with nondecreasing Lévy process input and constant release rate. Hence, for each k, the total content of the first k buffers has a generalized

Pollaczek–Khintchine distribution, so that we are able to provide a simple expression for the mean content at each node. For any Lévy input process, this enables us to provide simple necessary and sufficient conditions on the release rates for the average buffer contents to assume any prescribed vector of positive values. We also apply the continuous version of $L = \lambda W$ in Rolski and Stidham (1983) and Glynn and Whitt (1989) to calculate the long-run sojourn time in the network per particle.

By further analysis, we obtain additional partial characterizations of the steady-state joint distribution of the contents of the buffers. By relating a single buffer in the network to a linear fluid model with random disruptions, which was analysed by Chen and Yao (1992) and Kella and Whitt (1992a), we determine the marginal distribution of steady-state buffer content for each buffer. By applying a martingale result from Kella and Whitt (1992b), we also derive the two-dimensional joint distribution of the content of buffer j and the total content of the first $j-1$ buffers, for $2 \leqslant j \leqslant n$. We thus determine the full joint distribution in the case $n = 2$, which is *not* of product form. (See Kelly, 1979, and Walrand, 1988, for background on queueing networks with product-form steady-state distributions.)

The rest of this paper is organized as follows. In Section 7.2 we define the tandem fluid model with general input and prove that, consistent with intuition, the total content in the first k nodes has the simple one-dimensional form. In Section 7.3 we consider the tandem fluid model with non-decreasing Lévy process input. In Section 7.4 we obtain our partial characterizations of the joint distribution of the buffer contents. Finally, in Section 7.5 we show that the tandem fluid model can be represented as the limit of a sequence of conventional open queueing networks.

7.2 THE TANDEM FLUID MODEL

Consider n unlimited-capacity buffers in series connected by pipes. Let buffer j be the jth buffer visited. The flow rate (volume/time) through each pipe is determined by the flow rate of each particle and the cross-sectional area of the pipe. Let the pipe from buffer j to buffer $j+1$ be of length L_j with cross-sectional area A_j. Let the flow rate per particle on the pipe out of buffer j be r_j/A_j, so that the fluid can flow out of buffer j at rate r_j. Hence, if at time 0 the system is empty and a quantity x of fluid is put in buffer 1, then fluid enters buffer 2 at rate r_1 beginning at time A_1L_1/r_1. If no further input is made to buffer 1, then it becomes empty at time x/r_1, and flow into buffer 2 ceases at time $(A_1L_1 + x)/r_1$. The time lag $\tau_j \equiv A_jL_j/r_j$ is the propagation delay associated with the flow from buffer j to buffer $j+1$ through the connecting pipe.

If $r_j \leqslant r_{j+1}$ for some j, then the flow rate out of buffer $j+1$ is always greater

than or equal to the flow rate in, so that the buffer never fills. Hence, it suffices to assume that $r_1 > r_2 > \cdots > r_n$, and we do. In a communication network application, the flow rates might be the same on every link of the communication path. Then all buffering occurs at the initial access node, and our analysis has nothing to contribute. However, it is natural to consider unequal flow rates on the links.

Let the amount of fluid to arrive at the first buffer in the time interval $[0, t]$ be $X(t)$, where $X \equiv \{X(t)\colon t \geqslant 0\}$ is a real-valued stochastic process with non-decreasing sample paths that are right-continuous with left limits. For simplicity, assume that the system is initially empty. Let $W_j(t)$ be the buffer content (or work load) at buffer j at time t. Then

$$W_1(t) = X(t) - r_1 t + r_1 I_1(t) \tag{2.1}$$

and

$$\begin{aligned} &W_j(t + \tau_1 + \cdots + \tau_j) - W_j(\tau_1 + \cdots + \tau_{j-1}) \\ &\quad = W_j(t + \tau_1 + \cdots + \tau_j) \\ &\quad = r_{j-1}[t - I_{j-1}(t)] - r_j[t - I_j(t)], \qquad 2 \leqslant j \leqslant n, \end{aligned} \tag{2.2}$$

where

$$r_1 I_1(t) = -\inf_{0 \leqslant s \leqslant t} \{X(s) - r_1 s] \tag{2.3}$$

and

$$r_j I_j(t) = -\inf_{0 \leqslant s \leqslant t} \{r_{j-1}[s - I_{j-1}(s)] - r_j s\}, \qquad 2 \leqslant j \leqslant n. \tag{2.4}$$

Formulas (2.1)–(2.4) can be taken as definitions or derived from other first principles associated with the one-dimensional reflection map; see p. 4 of Beneŝ (1963), p. 19 of Harrison (1985) and p. 73 of Prabhu (1980).

From (2.1)–(2.4), we can determine the amount of fluid in each pipe at any time. Let $Y_j(t)$ represent the content of pipe j at time t. Then

$$\begin{aligned} Y_j(t + \tau_1 + \cdots + \tau_j) &= [r_j(t + \tau_1 + \cdots + \tau_j) - r_j I_j(t + \tau_1 + \cdots + \tau_j)] \\ &\qquad - [r_j(t + \tau_1 + \cdots + \tau_{j-1}) - I_j(t + \tau_1 + \cdots + \tau_{j-1})] \\ &= r_j \tau_j - r_j[I_j(t + \tau_1 + \cdots + \tau_j) \\ &\qquad - I_j(t + \tau_1 + \cdots + \tau_{j-1})], \qquad 1 \leqslant j \leqslant n. \end{aligned} \tag{2.5}$$

As mentioned earlier, it is convenient to focus on the partial sums

$$Z_j(t) = W_1(t) + \cdots + W_j(t + \tau_1 + \cdots + \tau_{j-1}), \qquad 1 \leqslant j \leqslant N. \tag{2.6}$$

A key property is that Z_j has the same one-dimensional reflection form as $Z_1 \equiv W_1$ for all j using r_j instead of r_1. This may be considered intuitively obvious, but we give a proof.

Theorem 2.1: For each j,

$$Z_j(t) = X(t) - r_j t + r_j I_j(t), \qquad t \geqslant 0, \tag{2.7}$$

where

$$r_j I_j(t) = -\inf_{0 \leqslant s \leqslant t} \{X(s) - r_j s\}, \qquad t \geqslant 0. \tag{2.8}$$

Proof. Formula (2.7) follows by adding the components in (2.1) and (2.2). Formula (2.8) is valid for $j = 1$ by (2.3). The other cases are established by induction. From (2.4) plus induction,

$$r_j I_j(t) = -\inf_{0 \leqslant s \leqslant t} \left\{ r_{j-1} s + \inf_{0 \leqslant u \leqslant s} \{X(u) - r_{j-1} u\} - r_j s \right\}, \qquad t \geqslant 0. \tag{2.9}$$

First,

$$\inf_{0 \leqslant u \leqslant s} \{X(u) - r_{j-1} u\} \leqslant X(s) - r_{j-1} s, \qquad s \geqslant 0,$$

so that

$$r_j I_j(t) \geqslant -\inf_{0 \leqslant s \leqslant t} \{r_{j-1} s + (X(s) - r_{j-1} s) - r_j s\} = -\inf_{0 \leqslant s \leqslant t} \{X(s) - r_j s\}.$$

We now show that the outer infimum in (2.9) can only be attained at an s or left limit $s-$ for which the inner infimum $\inf_{0 \leqslant u \leqslant s} \{X(u) - r_{j-1} u\}$ is attained at s or $s-$. Suppose that the inner infimum is attained at $u^*(s)$ or $u^*(s)-$ with $u^*(s) < s$. Then, since $r_{j-1} > r_j$,

$$\begin{aligned} r_{j-1} s + \inf_{0 \leqslant u \leqslant s} \{X(u) - r_{j-1} u\} - r_j(s) \\ &= r_{j-1} s + \inf_{0 \leqslant u \leqslant u^*(s)} \{x(u) - r_{j-1} u\} - r_j s \\ &> r_j u^*(s) + \inf_{0 \leqslant u \leqslant u^*(s)} \{X(u) - r_{j-1} u\} - r_j u^*(s). \end{aligned}$$

Hence,

$$r_j I_j(t) \leqslant -\inf_{0 \leqslant s \leqslant t} \{X(s) - r_j s\}$$

and the proof is complete. ■

Remark 2.2: If $X(t)$ is a pure jump process, then $I_j(t)$ depicts the cumulative time in $[0, t]$ that the first j buffers are simultaneously empty.

In communication networks we are interested in the sojourn time of messages, i.e. the time from the arrival of the first packet of the message at the first buffer until the departure of the last packet from the network. In our model, we represent a message as a batch arrival. It is significant that the sojourn time of a batch is easily related to the sojourn time of the first particle in a batch. The following is an elementary consequence of the fluid flow.

Proposition 3.2: The sojourn time of an arriving batch of size x is equal to the sojourn time of the first particle in the batch plus x/r_n.

7.3 THE TANDEM FLUID MODEL WITH LÉVY INPUT

In order to obtain tractable expressions for the steady-state distribution, we now assume that the input process X is a non-decreasing Lévy process. (See Chapter 3 of Prabhu, 1980.) In particular, X is defined on an underlying probability space $(\Omega, \mathscr{F}, P\}$ endowed with a standard filtration $\{\mathscr{F}_t: t \geqslant 0\}$. We assume that $X(0) = 0, X(t)$ is adapted to $\mathscr{F}_t$ and $X(u) - X(t)$ is independent of $\mathscr{F}_t$ and distributed as $X(u - t)$ for $0 \leqslant t < u$, where $X(t) \geq 0$ with probability 1. We assume that $X(t)$ has Laplace transform

$$E\,\mathrm{e}^{-\alpha X(t)} = \mathrm{e}^{-\phi(\alpha)t}, \qquad t \geqslant 0, \tag{3.1}$$

where $\phi(\alpha)$ is the characteristic exponent.

Note that $(Z_1(t), \ldots, Z_n(t))$ defined in (2.6) is also adapted to $\mathscr{F}_t$ for each t. As an easy consequence of Theorem 4, p. 78, of Prabhu (1980) and Theorem 2.1, we see that under minor regularity conditions $Z_j(t)$ has a generalized Pollaczek–Khintchine distribution for each j. (Also see Gani and Pyke, 1960, and Chapter 7 of Prabhu, 1965.)

Theorem 3.1: (*a*) If the input process X is a non-decreasing Lévy process and $\rho \equiv EX(1) < r_j$, then $Z_j(t)$ converges in distribution to a proper limit $Z_j(\infty)$ as $t \to \infty$ for each j and

$$E\,\mathrm{e}^{-\alpha Z_j(\infty)} = \frac{\alpha(r_j - \rho)}{\alpha r_j - \phi(\alpha)} \qquad \text{for} \quad \alpha > 0. \tag{3.2}$$

(*b*) If, in addition, $\sigma^2 \equiv \operatorname{Var} X(1) < \infty$, then

$$EZ_j(\infty) = \frac{\sigma^2}{2(r_j - \rho)}. \tag{3.3}$$

■

Example 3.2: If X is a compound Poisson process with Poisson rate λ and i.i.d. random jumps $J_n, n \geqslant 1$, having mean m and variance $\hat{\sigma}^2$, then $\phi(\alpha) = \lambda(1 - \psi(\alpha))$ where $\psi(\alpha) = E(e^{-\alpha J_1})$, $\rho \equiv EX(1) = \lambda m$ and $\sigma^2 \equiv \operatorname{Var} X(1) = \lambda(\hat{\sigma}^2 + m^2)$. Moreover, in this case it is easy to see that $\{Z_j(t): t \geqslant 0\}$, $\{V_{\lambda r_j^{-1}, J}(r_j t): t \geqslant 0\}$ and $\{r_j V_{\lambda, J r_j^{-1}}(t): t \geqslant 0\}$ all have the same finite-dimensional distributions, where $\{V_{\lambda, J}(t): t \geqslant 0\}$ is the $M/G/1$ virtual waiting time process with arrival rate λ and generic service time J.

We can apply Theorem 3.1 to determine, for any given Lévy process $X(t)$, necessary and sufficient conditions on the rates r_j for the mean buffer contents to assume any prescribed values. In particular, there is one and only one vector of release rates yielding each vector of expected buffer contents.

Corollary 3.3: Under the assumptions of Theorem 3.1,

$$EW_j(\infty) = \frac{\sigma^2(r_{j-1} - r_j)}{2(r_j - \rho)(r_{j-1} - \rho)}, \tag{3.4}$$

so that $EW_j(\infty) = x_j > 0$ for all j if and only if

$$r_j = \rho + \frac{\sigma^2}{2(x_1 + \cdots + x_j)}, \qquad 1 \leqslant j \leqslant n. \tag{3.5}$$

Proof. Apply (3.3). ■

We can also apply Theorem 3.1 to determine the average sojourn time in the entire network, using a continuous version of $L = \lambda W$; see Rolski and Stidham (1983) and Glynn and Whitt (1989).

Theorem 3.4: Under the assumptions of Theorem 3.1(b), the average sojourn time in the network per particle converges almost surely to

$$\sum_{j=1}^{n} \tau_j + \frac{\sigma^2}{2\rho(r_n - \rho)}. \tag{3.6}$$

Proof. We apply Theorem 6 of Glynn and Whitt (1989). In the framework of Glynn and Whitt (1989), $X(t)$ plays the role of $S_1(t)$ in (3) there. Its inverse $T_1(s)$, defined by

$$T_1(s) = \inf\{t \geqslant 0: S_1(t) > s\}, \qquad s \geqslant 0, \tag{3.7}$$

gives the time of arrival associated with particle s. (The particles are indexed by points on the real line.) The long-run average sojourn time is defined as

$$\tau = \lim_{u \to \infty} \frac{1}{u} \int_0^u w(s)\, ds, \tag{3.8}$$

where $w(s)$ is the time spent in the system by particle s; see Sections 1, 1.6 and 4.4 of Glynn and Whitt (1989). The limit in (3.6) is identical to the long-run average per unit time, defined by

$$\lim_{t\to\infty} \frac{1}{X(t)} \int_0^{X(t)} w(s)\,\mathrm{d}s. \tag{3.9}$$

In order to establish the a.s. limit for the time-average buffer content, we use regenerative structure. The regenerative cycles will be associated with the process $\{(Z_1(t), \ldots, Z_n(t)) \colon t \geqslant 0\}$. The cycles begin with $Z_1(0) = \cdots = Z_n(0) = 0$. For any $x > 0$, the cycle ends at time

$$T_x = \inf\{t \geqslant 0 \colon x + X(t) - r_n t = 0\} = \inf\{t \geqslant 0 \colon I_n(t) = x\}. \tag{3.10}$$

By p. 79 of Prabhu (1980), $ET_x = x/(r_n - \rho) < \infty$. This, together with Theorem 3.1(b), implies that the time-average work load in the n buffers converges a.s. to $\sigma^2/2(r_n - \rho)$. The corresponding term in (3.6) is obtained by dividing by ρ, invoking $L = \lambda W$. The time spent by each particle in pipe j is τ_j, so that the time average converges with probability 1 trivially. It remains to verify the extra conditions in Theorem 6 of Glynn and Whitt (1989). First, $t^{-1}X(t) \to \rho$ a.s. as $t \to \infty$ by the strong law of large numbers, using the Lévy property and the moment condition $E[X(1)] < r_n$. Finally, to see that $s^{-1}w(s) \to 0$ a.s. as $s \to \infty$, note that the limit of $s^{-1}w(s)$ as $s \to \infty$ coincides with the limit $w(X(t))/X(t)$ as $t \to \infty$. Also, $w(X(t))$ is dominated above by $\sum_{j=1}^{n} (\tau_j)$ plus the maximum cycle length of all the cycles up to time t, including the one covering t. Let T_j be the jth cycle. Since $EX(1)^2 < \infty$, $ET_1^2 < \infty$; see p. 79 of Prabhu (1980). By Chebyshev, for any $\varepsilon > 0$ and n,

$$P(T_n > n\varepsilon) \leqslant \frac{E(T_n^2)}{n^2\varepsilon}.$$

By Borel–Cantelli, since

$$\sum_{n=1}^{\infty} P(T_n > n\varepsilon) < \infty,$$

$$P(n^{-1}T_n > \varepsilon \text{ infinitely often}) = 0 \quad \text{a.s.},$$

so that

$$n^{-1}\max\{T_1, \ldots, T_n\} \to 0 \text{ as } n \to \infty \quad \text{a.s.}$$

Let $N(t)$ count the number of cycles in $[0, t]$. By the above

$$\frac{w(X(t))}{X(t)} \leqslant \frac{1}{X(t)} \sum_{j=1}^{n} \tau_j + \frac{N(t)+1}{X(t)} \frac{1}{N(t)+1} \max\{T_1, \ldots, T_{N(t)+1}\}$$

which converges to 0 as $t \to \infty$ w.p.1. ■

Corollary 3.5: If, in addition to the assumptions of Theorem 3.4, the Lévy process X is compound Poisson with jumps having an exponential or a geometric distribution, then the average sojourn time in the network for the batches converges almost surely to the quantity in (3.6).

Proof. For the exponential jumps, apply PASTA (Poisson Arrivals See Time Averages) (see Wolff, 1982, or Melamed and Whitt, 1990), letting 'time' be the particle index, so that the embedded sequence corresponding to the last particle in a jump is a Poisson process. This argument is the continuous analogue of the argument in Halfin (1983), who observed that the discrete-time analogue of PASTA applied with geometrically distributed jumps. See Whitt (1983) and Makowski *et al.* (1989) for more on the discrete-time analogue of PASTA. This argument shows that the average sojourn time of the last particle in a jump is the same as the average sojourn time of an arbitrary particle when the jumps have an exponential or a geometric distribution. ■

We can apply $L=\lambda W$ the other way to determine the long-run time-average content of the pipes.

Proposition 3.6: Under the assumptions of Theorem 3.4, the long-run average content of pipe j is $\rho\tau_j$.

Proof. Apply $L = \lambda W$ again using the pipe j term τ_j in (3.6). ■

7.4 MORE ABOUT THE STEADY-STATE DISTRIBUTION

In this section we derive the limiting distributions of $W_j(\tau_1 + \cdots + \tau_j + t)$ and $(Z_{j-1}(t), W_j(\tau_1 + \cdots + \tau_j + t))$ as $t \to \infty$ for each j. Without loss of generality and for the sake of simplified notation, we will assume (in this section only) that there are no pipes, i.e. the flow from the jth buffer to the $(j + 1)$st occurs instantaneously at a rate of r_j so that $\tau_1 = \cdots = \tau_n = 0$. We will also assume that the input process to the buffer is compound Poisson with exponent $\phi(\alpha) = \lambda(1 - \psi(\alpha))$ as in Example 3.2. Let $W_j(\infty)$ be a random variable whose distribution is the limiting distribution of $W_j(t)$. Since the entire process is regenerative, as observed in Section 7.3 (see eqn (3.9)), and the cycle lengths have a non-arithmetic distribution, the existence of the limiting distribution is assured.

The important observation is to note that $W_j(t)$ is increasing at a rate of $r_{j-1} - r_j$ whenever $Z_{j-1}(t) > 0$ (equivalently $W_{j-1}(t) > 0$). Also, if we denote by D_{ji}, $U_{j,i-1}$ the length of the ith interval during which $Z_{j-1} > 0$, $Z_{j-1} = 0$, respectively, then we see that $\{(D_{ji}, U_{ji}): i \geqslant 1\}$ is a sequence of i.i.d. random vectors. To start a regeneration point, we start observing $W_j(\cdot)$ right after the first batch arrives, rather than from time zero. This clearly does not change the limiting distribution and is done in order to readily use available results. Let $u_j = EU_{j1}$ and $d_j = ED_{j1}$. We know from assumptions in previous sections that $r_j u_j > (r_{j-1} - r_j)d_j$. Hence, the process $\{W_j(t): t \geqslant 0\}$ coincides with the content process in the linear fluid model with random disruptions introduced by Chen and Yao (1992) and further analysed by Kella and Whitt (1992a). In Kella and Whitt (1992a) we observed that there is a direct connection between this process and the work load or virtual waiting time process in the $GI/G/1$ queue with (in this case) inter-arrival times $r_j U_{ji}$ and service times $(r_{j-1} - r_j)D_{ji}$. More precisely, if we denote by V_j a random variable with the steady-state distribution of the work load in the $GI/G/1$ queue, then $(W_j(\infty) \mid W_j(\infty) > 0)$ has the same distribution as $(V_j \mid V_j > 0)$. For the corresponding $GI/G/1$ queue, it is well known that

$$P(V_j > 0) = (r_{j-1} - r_j)d_j/r_j u_j$$

and it is also easily shown that

$$P(W_j(\infty) > 0) = p_{uj}P(V_j > 0) + p_{dj},$$

where $p_{uj} = 1 - p_{dj} = u_j/(d_j + u_j)$. Hence,

$$P(W_j(\infty) > t) = \pi_j P(V_j > t), \qquad t \geqslant 0, \tag{4.1}$$

where

$$\pi_j = \frac{P(W_j(\infty) > 0)}{P(V_j > 0)} = \left(\frac{u_j}{u_j + d_j}\right)\left(\frac{r_{j-1}}{r_{j-1} - r_j}\right) > 1. \tag{4.2}$$

In our case, D_j is distributed as the busy period in an $M/G/1$ queue with arrival rate λ and service times with Laplace–Stieltjes transform (LST) $\psi(\alpha/r_{j-1})$, since $Z_{j-1}(t)/r_{j-1}$ is the virtual waiting time process of a standard $M/G/1$ queue. This means that the LST of D_{j1}, denoted by $\hat{d}_j(\alpha)$ is the minimal positive root of the functional equation

$$\hat{d}_j(\alpha) = \psi[(\alpha + \lambda - \lambda\hat{d}_j(\alpha))/r_{j-1}]. \tag{4.3}$$

Consequently, $d_j = m/(r_{j-1} - \lambda m)$, where $m = -\psi'(0)$. As mentioned before, U_{j1} has an exponential distribution with parameter λ. This implies that the distribution of V_j is determined by the Pollaczek–Khintchine formula for an $M/G/1$ queue with arrival rate λ/r_j and service times with LST $\hat{d}_j((r_{j-1} - r_j)\alpha)$. Summarizing, we have the following result.

Theorem 4.1: If the external input to the first buffer is compound Poisson with exponent $\lambda(1 - \psi(\alpha))$ as in Example 3.2, with $m \equiv -\psi'(0)$ and

$\hat{\rho}_j \equiv \lambda m/r_j < 1$, then

$$F_j(t) \equiv P(W_j(\infty) \leqslant t) = 1 - \pi_j + \pi_j \sum_{i=0}^{\infty} (1 - \rho_j)\rho_j^i H_j^{*i}(t)$$

$$= 1 - \hat{\rho}_j + \hat{\rho}_j \sum_{i=1}^{\infty} (1 - \rho_j)\rho_j^{i-1} H_j^{*i}(t), \qquad (4.4)$$

where

$$\pi_j = \frac{r_{j-1} - r^j \hat{\rho}_j}{r_{j-1} - r_j} = \frac{r_{j-1} - \rho}{r_{j-1} - r_j}, \qquad \rho_j = \frac{(r_{j-1} - r_j)\hat{\rho}_j}{r_{j-1} - r_j\hat{\rho}_j} \qquad (4.5)$$

and H_j^{*i} is the ith fold convolution of H_j with LST

$$\hat{h}_j(\alpha) = \frac{1 - \hat{d}_j((r_{j-1} - r_j)\alpha)}{\alpha(r_{j-1} - r_j)d_j}, \qquad (4.6)$$

where $\hat{d}_j(\alpha)$ is the minimal positive root of the equation

$$\hat{d}_j(\alpha) = \psi[(\alpha + \lambda - \lambda\hat{d}_j(\alpha))/r_{j-1}] \qquad (4.7)$$

and $d_j = m/(r_{j-1} - \lambda m)$. ■

From Theorem 4.1 we can calculate the moments of $W_j(\infty)$. Combining the variance of $W_j(\infty)$ with the variance of $Z_j(\infty)$, we can calculate the covariance between $Z_{j-1}(\infty)$ and $W_j(\infty)$. We find that the correlation between $Z_{j-1}(\infty)$ and $W_j(\infty)$, denoted $\text{cor}(Z_{j-1}(\infty), W_j(\infty))$, can assume any value in the interval $(0, 1/\sqrt{3})$.

Corollary 4.2: Under the assumptions of Theorem 4.1,

$$EW_j(\infty) = \frac{EZ_j(\infty)}{\pi_j} \qquad (4.8)$$

and

$$\text{Var } W_j(\infty) = \frac{\text{Var } Z_j(\infty) + 2EZ_{j-1}(\infty)EZ_j(\infty)}{\pi_j^2} \qquad (4.9)$$

with π_j in (4.5), $EZ_j(\infty)$ and $EZ_{j-1}(\infty)$ in (3.3), so that

$$\text{Var } Z_j(\infty) = \left[\frac{\rho}{r_j - \rho}\frac{EJ^2}{2EJ}\right]^2 + \frac{\rho}{r_j - \rho}\frac{EJ^3}{3EJ}, \qquad (4.10)$$

and

$$\operatorname{cor}(Z_{j-1}(\infty), W_j(\infty)) = \frac{1}{2}\sqrt{\frac{pc^2}{c^2 - 1 + ((1/p) - 1)(c - 1)}}$$
$$\leqslant \frac{1}{2}\sqrt{\frac{c^2}{c^2 - 1}} \leqslant \frac{1}{\sqrt{3}}, \tag{4.11}$$

where $p = (r_j - \rho)/(r_{j-1} - \rho) \in (0, 1)$ and

$$c = 2 + [(r_{j-1} - \rho)4\lambda EJ^3]/[3(EJ^2)^2] \in (2, \infty),$$

so that $\operatorname{cor}(Z_{j-1}(\infty), W_j(\infty))$ can assume any value in the interval $(0, 1/\sqrt{3})$. ■

With the aid of Theorem 4.1 we are now ready to obtain the limiting distribution of the pair $(Z_{j-1}(t), W_j(t))$, whose existence is assured from regenerative process theory.

Theorem 4.3: With the notation and assumptions of Theorem 4.1, let $(Z_{j-1}(\infty), W_j(\infty))$ denote a pair of random variables having the limiting distribution of $(Z_{j-1}(t), W_j(t))$, for $2 \leqslant j \leqslant n$. Then

$$E\,\mathrm{e}^{-(\alpha Z_{j-1}(\infty) + \beta W_j(\infty))} = \frac{(\alpha - \beta)(r_{j-1} - r_j)E\,\mathrm{e}^{-\beta W_j(\infty)} + (r_j - \lambda m)\alpha}{r_{j-1}\alpha - (r_{j-1} - r_j)\beta - \lambda(1 - \psi(\alpha))}, \tag{4.12}$$

where the distributions of $W_j(\infty)$ (from which the LST is immediate) is given in Theorem 4.1.

Proof. For this proof, without loss of generality, we may assume that $n = j$, and we do. We begin by defining

$$M(t) = [r_{j-1}\alpha - (r_{j-1} - r_j)\beta - \lambda(1 - \psi(\alpha))] \int_0^t \mathrm{e}^{-(\alpha Z_{j-1}(s) + \beta W_j(s))}\,\mathrm{d}s$$
$$+ 1 - \mathrm{e}^{-(\alpha Z_{j-1}(t) + \beta W_j(t))} - r_{j-1}(\alpha - \beta)\int_0^t \mathrm{e}^{-\beta W_j(s)}\,\mathrm{d}I_{j-1}(s) - \beta r_j I_j(t), \tag{4.13}$$

and noting that Theorem 2 of Kella and Whitt (1992b) implies that $\{M_t | t \geqslant 0\}$ is a zero-mean martingale. Let T_x be as in equation (3.9). Applying Doob's optional stopping theorem to the martingale $\{M_t | t \geqslant 0\}$, with respect to the bounded stopping time $T_x \wedge t \equiv \min(T_x, t)$, gives $EM_{T_x \wedge t} = 0$. Before proceeding we first make the observations that $Z_{j-1}(T_x) = W_j(T_x) = 0$, that $\mathrm{d}I_{j-1}(s) = 1_{\{Z_{j-1}(s) = 0\}}\,\mathrm{d}s$, that (by definition) $r_j I_j(T_x) = x$ and that

$ET_x = x/(r_j - \lambda m)$. Since all of the components of M_t are either monotone or bounded functions of t, then by monotone and bounded convergence, applied to each component separately, we obtain

$$EM_{T_x} = \lim_{t \to \infty} EM_{T_x \wedge t} = 0. \tag{4.14}$$

Upon dividing by ET_x and rearranging terms, we obtain

$$\begin{aligned}[][r_{j-1}\alpha - (r_{j-1} - r_j)\beta - \lambda(1 - \psi(\alpha))] \frac{1}{ET_x} E \int_0^{T_x} \mathrm{e}^{-(\alpha Z_{j-1}(s) + \beta W_j(s))} \, \mathrm{d}s \\ = r_{j-1}(\alpha - \beta) \frac{1}{ET_x} E \int_0^{T_x} \mathrm{e}^{-\beta W_j(s)} 1_{\{Z_{j-1}(s)=0\}} \, \mathrm{d}s + \beta(r_j - \lambda m). \end{aligned} \tag{4.15}$$

Equation (4.15), together with regenerative process theory and the fact that $P[Z_{j-1}(\infty) = 0] = 1 - \lambda m / r_{j-1}$, gives

$$\begin{aligned} &E\,\mathrm{e}^{-(\alpha Z_{j-1}(\infty) + \beta W_j(\infty))} \\ &= \frac{(\alpha - \beta)(r_{j-1} - \lambda m) E[\mathrm{e}^{-\beta W_j(\infty)} | Z_{j-1}(\infty) = 0] + (r_j - \lambda m)\beta}{r_{j-1}\alpha - (r_{j-1} - r_j)\beta - \lambda(1 - \psi(\alpha))}. \end{aligned} \tag{4.16}$$

Finally the proof is complete if we observe that

$$\begin{aligned} &(r_{j-1} - \lambda m) E[\mathrm{e}^{-\beta W_j(\infty)} | Z_{j-1}(\infty) = 0] \\ &\qquad = (r_{j-1} - r_j) E\,\mathrm{e}^{-\beta W_j(\infty)} + (r_j - \lambda m), \end{aligned} \tag{4.17}$$

as can be seen by setting $\alpha = 0$ in eqn (4.16). ■

7.5 CONVERGENCE OF QUEUEING NETWORKS

In this final section we indicate how our network fluid model can be represented as the limit of a sequence of conventional queueing network models. We choose a limiting regime that seems relevant for high-speed communication networks. In particular, we let the conventional networks have batch Poisson arrival processes with the batch sizes growing and the service times decreasing. This is intended to represent longer messages being transported at higher speeds. We could also consider a more elaborate limit in which the individual arrivals (packets) in a batch (message) are separated by small spaces instead of arriving at one instant, but such that the spacing between arrivals within a batch is asymptotically negligible in the limit, but we do not treat this modification (see Remark 5.3 below).

We define a sequence of conventional models indexed by k. For each k, let there be $2n - 1$ queues with unlimited capacity in series, with the odd-numbered queues having one server and the even-numbered queues having infinitely many servers. The even-numbered queues are pure-delay nodes intended to represent the connecting pipes. For simplicity, we specify the service times for all the models and all the queues with a single sequence of non-negative random variables $(v_m: m \geqslant 1\}$. The service time of customer m at queue $(2j-1)$ of model k is $v_m/r_j k$; the service time of customer m at queue $2j$ of model k is τ_j. (The service time of customer m is determined by the service requirement v_m (e.g. packet length) and the service rate prevailing at the node. We think of the service rate in model k being proportional to k but the lengths of the pipes (distances between nodes) being proportional to k too, so that the time spent in the pipe does not change with k.)

Let the arrival processes be determined by a Poisson counting process $\{A(t): t \geqslant 0\}$ with rate λ and an i.i.d. sequence of batch sizes $\{B_i: i \geqslant 1\}$ with $EB_1 = m$ and Var $B_1 = \hat{\sigma}^2 < \infty$. In model k, let the arrival process be a batch Poisson process with Poisson counting process $A(t)$, independent of k, and let the ith batch be kB_i.

Let $X_k(t)$ represent a *scaled* total input of service requirement to the kth model in [0, t]; i.e. let

$$X_k(t) = \sum_{i=1}^{A(t)} J_{ki}, \qquad t \geqslant 0, \tag{5.1}$$

where $A(t)$ is a Poisson counting process having rate α,

$$J_{ki} = \sum_{m=kS_{i-1}+1}^{m=kS_i} (v_m/k), \qquad i \geqslant 1, \tag{5.2}$$

and

$$S_i = B_1 + \cdots + B_i, \qquad i \geqslant 1. \tag{5.3}$$

What we want now is for $(J_{k1}, \ldots, J_{ki})$ to converge w.p.1 to $(B_1 Ev_1, \ldots, B_i Ev_1)$ for each i as $k \to \infty$. Then $\{X_k(t): t \geqslant 0\}$ converges to $\{X(t): t \geqslant 0\}$ uniformly on compact time intervals (u.o.c.) with probability 1, where X is a compound Poisson process with an i.i.d. jump sequence $\{B_i Ev_1: i \geqslant 1\}$. Of course, a sufficient condition on the service times is for them to be i.i.d. with $Ev_1 < \infty$.

Let $W_{kj}(t)$ be the sum of all service times of customers currently in queue $(2j-1)$ at time t, let $I_{kj}(t)$ be the cumulative idle time in queue $(2j-1)$ in the interval $[0, t]$, and let $Y_{kj}(t)$ represent the sum of all service times of all customers in queue $2j$ at time t, all for model k. For any k, $W_{kj}(t)$, $I_{kj}(t)$, and $Y_{kj}(t)$ do not quite satisfy the relations in Section 7.2 because here the customers flow discretely instead of continuously as a fluid. However, because of the scaling of the service times, the vector of these processes converges uniformly on compact time intervals w.p.1 to the vector process

$$[W_1(t), \ldots, W_n(t), I_1(t), \ldots, I_n(t), Y_1(t), \ldots, Y_{n-1}(t)]$$

associated with the fluid model in Section 7.2, determined by the limiting compound Poisson process X and the parameters r_j and τ_j.

Summarizing, we have the following result.

Theorem 5.1: If $(J_{k1}, \ldots, J_{ki}) \rightarrow (B_1 Ev_1, \ldots, B_i Ev_1)$ w.p.1 as $k \rightarrow \infty$ for every i, then

$$\{X_k(t): t \geqslant 0\} \rightarrow \{X(t): t \geqslant 0\} \qquad \text{u.o.c.} \quad \text{w.p.1 as } k \rightarrow \infty$$

and

$$\{(W_{kj}(t), I_{kj}(t), Y_{kj}(t), 1 \leqslant j \leqslant n): t \geqslant 0\} \rightarrow \{(W_j(t), I_j(t), Y_j(t), 1 \leqslant j \leqslant n): t \geqslant 0\} \qquad \text{u.o.c.} \quad \text{w.p.1 as } k \rightarrow \infty,$$

where the limit is associated with the tandem fluid model associated with the compound Poisson process $\{X(t): t \geqslant 0\}$.

Proof. If the kth queueing model acted as a fluid model with continuous linear flow out of each buffer, then the convergence would be a simple consequence of the continuity of the multidimensional reflection map in the topology of uniform convergence on finite time intervals (Harrison and Reiman, 1981; Chen and Mandelbaum, 1991; Chen and Whitt, 1991), which in the special case of a tandem network is easy to establish directly from the continuity of the one-dimensional reflection map. Hence, it suffices to show that the kth queueing model in which customers are treated discretely gets arbitrarily close to the kth queueing model treated as a fluid model as $k \rightarrow \infty$. For each k, the two models under consideration have the same external arrival process and service requirements, but for the queueing model the flow involves discrete customers rather than fluid. The two models get suitably close because the service times are being divided by k as $k \rightarrow \infty$.

In particular, with k fixed, the two models under consideration have the same arrival process at the first queue. It is easy to see that the last particle of each customer starts and completes service at the first queue at the same time in both models. Hence, the maximum difference in the work loads over all times in the interval $[0, t]$ is bounded above by the maximum service time of all the arrivals in $[0, t]$, say $m_1(t)$. Let $X_{1j}(t)$ and $X_{2j}(t)$ represent the total input of service requirement to buffer j in time interval $[0, t]$ for the fluid and discrete models, respectively. By above, $X_{11}(t) = X_{21}(t)$ for all $t \geqslant 0$ and

$$X_{12}(s) \geqslant X_{22}(s) \geqslant X_{12}(s) - m_1(t), \qquad 0 \leqslant s \leqslant t.$$

By induction,

$$X_{1j}(s) \geqslant X_{2j}(s) \geqslant X_{1j}(s) - \sum_{i=1}^{j-1} m_i(t), \qquad 0 \leqslant s \leqslant t,$$

for each $j \geqslant 1$, where $m_i(t)$ is the maximum service time of all customers at buffer i among all arrivals to buffer 1 in $[0, t]$. It suffices to show that $m_{ki}(t) \to 0$ w.p.1 as $k \to \infty$ for each i, where $m_{ki}(t)$ is $m_i(t)$ in model k, but this follows from the assumed convergence $(J_{k1}, \ldots, J_{kl}) \to (B_1 Ev_1, \ldots, B_l Ev_l)$ with probability 1. (Note that $\max\{v_i: 1 \leqslant i \leqslant mk)/k \to 0$ with probability 1 as $k \to \infty$ whenever $k^{-1} \sum_{i=1}^{mk} v_i \to c$ w.p.1 as $k \to \infty$.) ■

Remark 5.2: Our special construction has enabled us to obtain w.p.1 convergence in Theorem 5.1. If we only assumed that, for each k, $\{A_k(t): t \geqslant 0\}$ is a stochastic counting process that converges in distribution as $k \to \infty$ to the Poisson process $\{A(t): t \geqslant 0\}$ in $D[0, \infty)$ with Skorohod's (1956) J_1 topology, then the conclusion of Theorem 5.1 would be convergence in distribution in $D[0, \infty)$ with Skorohod's (1956) J_1 topology.

Remark 5.3: If we introduce spacing between the arrivals within a batch that is asymptotically negligible, then we would need to use convergence in Skorohod's (1956) M_1 topology on the function space $D(0, \infty)$ as in Chen and Whitt (1991).

REFERENCES

Anick, D., Mitra, D., and Sondhi, M. M. (1982). Stochastic theory of a data handling system. *Bell System Tech. J.*, **61**, 1871–94.

Beneŝ, V. (1963). *General stochastic processes in the theory of queues.* Addison-Wesley, Reading, Mass.

Chen, H. and Mandelbaum, A. (1991). Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Operat. Res.*, **16**, 408–46.

Chen, H. and Whitt, W. (1991). Diffusion approximations for queueing networks with service interruptions, submitted for publication.

Chen, H. and Yao, D. D. (1992). A fluid model for systems with random disruptions. *Operat. Res.*, **40**, Supplement 2, S239–47.

Gani, J. and Pyke, R. (1960). The content of a dam as the supremum of an infinitely divisible process. *J. Math. and Mech.*, **9**, 639–52.

Glynn, P. W. and Whitt, W. (1989). Extensions of the queueing relations $L = \lambda W$ and $H = \lambda G$. *Operat. Res.*, **37**, 634–44.

Halfin, S. (1983). Batch delays versus customer delays. *Bell System Tech. J.*, **62**, 2011–15.

Harrison, J. M. (1985). *Brownian motion and stochastic flow systems.* Wiley, New York.

Harrison, J. M. and Reiman, M. I. (1981). Reflected Brownian motion on an orthant. *Ann. Probab.*, **9**, 302–8

Kella, O. and Whitt, W. (1992a). A storage model with a two-state random environment, *Operat. Res.*, **40**, Supplement 2, S257–62.

Kella, O. and Whitt, W. (1992b). Useful martingales for stochastic storage processes with Lévy input, *J. Appl. Prob.*, **29**, June.

Kelly, F. P. (1979). *Reversibility and stochastic networks*. Wiley, Chichester.

Makowski, A., Melamed, B., and Whitt, W. (1989). On averages seen by arrivals in discrete time. *Proc. 28th IEEE Conference on Decision and Control*, 1084–6.

Melamed, B. and Whitt, W. (1990). On arrivals that see time averages. *Operat. Res.*, **38**, 156–72.

Mitra, D. (1988). Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Adv. Appl. Prob.*, **20**, 646–76.

Newell, G. F. (1982). *Applications of queueing theory* (2nd edn). Chapman and Hall, London.

Prabhu, N. U. (1965). *Queues and inventories*. Wiley, New York.

Prabhu, N. U. (1980). *Stochastic storage processes*. Springer-Verlag, New York.

Rolski, T. and Stidham, S. Jr. (1983). Continuous versions of the queueing formulas $L = \lambda W$ and $H = \lambda G$. *Operat. Res. Letters*, **18**, 211–15.

Skorohod, A. V. (1956). Limit theorems for stochastic processes. *Theor. Prob. Appl.*, **1**, 261–90.

Walrand, J. (1988). *An introduction to queueing networks*. Prentice-Hall, Englewood Cliffs.

Whitt, W. (1983). Comparing batch delays and customer delays. *Bell System Tech. J.*, **62**, 2001–9.

Wolff, R. W. (1982). Poisson arrivals see time averages. *Operat. Res.*, **30**, 223–31.