Linear Stochastic Fluid Networks

by

Offer Kella [1] *and Ward Whitt* [2]

April 17, 1997

Revised: February 17, 1998

[1]Department of Statistics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel; mskella@olive.mscc.huji.ac.il

[2]Room A117, AT&T Laboratories, 180 Park Avenue, Building 103, Florham Park, NJ 07932-0971, USA; wow@research.att.com

Abstract

We introduce open stochastic fluid networks that can be regarded as continuous analogs or fluid limits of open networks of infinite-server queues. Random exogenous input may come to any of the queues. At each queue, a cdf-valued stochastic process governs the proportion of the input processed by a given time after arrival. The routing may be deterministic (a specified sequence of successive queue visits) or proportional, i.e., a stochastic transition matrix may govern the proportion of output routed from one queue to another. This stochastic fluid network with deterministic cdf's governing processing at the queues arises as the limit of normalized networks of infinite-server queues with batch arrival processes where the batch sizes grow. In this limit, one can think of each particle having an evolution through the network, depending on its time and place of arrival, but otherwise independent of all other particles. A key property associated with this independence is the linearity: The workload associated with a superposition of inputs, each possibly having its own pattern of flow through the network, is simply the sum of the component workloads. Just like infinite-server queueing models, the tractability makes the linear stochastic fluid network a natural candidate for approximations.

Keywords: fluid models, stochastic fluid networks, storage networks, linear stochastic fluid networks, Lévy processes, infinite-server queues, queueing networks, shot noise, multidimensional shot noise

1. Introduction

This paper is a continuation of our study of stochastic fluid networks begun in Kella and Whitt [25] and continued in Kella [22, 23, 24], Kella and Whitt [26] and references therein. The basic stochastic fluid network has random external inputs, but all internal flows are deterministic and continuous like fluid (occurring at a deterministic rate, conditional on the system state). Stochastic fluid networks are natural models for stochastic storage networks in which the random fluctuations within the network occur in a shorter time scale than the random fluctuations in the external inputs. A possible application is a communication network composed of high-speed switches serving randomly arriving large messages which are broken up into fixed-length packets or "cells" and sent from origin to destination. The fluid property of the internal flows provides a simplification that enables us to say more about model behavior. For example, for certain cases (see [25, 22, 23]) we derived the steady-state distribution of the contents in a two-buffer stochastic fluid network with Lévy external input process (which is never product form), as well as the steady state covariance matrix in the multi-buffer tandem case (see [22]).

In this paper we introduce a new stochastic fluid model that is even more tractable. It is a linear stochastic fluid model, in which the movement of separate particles, conditional on the time and place they enter the network, can be thought of as being mutually independent. However, since there are in general uncountably infinitely many particles under consideration, we do not directly define stochastic behavior of individual particles. Instead, we specify what happens to deterministic proportions of the fluid after it arrives.

The buffer content vector $[W_1(t), \ldots, W_m(t)]$ associated with a stochastic fluid network having m queues can be regarded as the limit as $n \to \infty$ of the scaled number of busy servers at each queue, $n^{-1}(Q_1^{(n)}(t), \ldots, Q_m^{(n)}(t))$, in a network of m infinite-server queues with batch arrivals, denoted by $(G^X/G/\infty)^m$, where the batch sizes in model n are the batch sizes in model 1 multiplied by n. In the $(G^X/G/\infty)^m$ queueing model, arrivals in the same batch have i.i.d. stochastic paths through the network. In contrast, in the limiting fluid model, by virtue of the law of large numbers, deterministic proportions follow those same routes. Similarly, instead of i.i.d. service times at a queue governed by a cdf G, a proportion $G(t)$ of all fluid arriving at a given time completes processing and departs in the interval $[0, t]$ after arrival. (For related limits, see Kurtz [28].) We actually consider a generalization of the fluid model specified so far, in which the cdf G is allowed to be a stochastic process (see Section 2).

Thus the linear stochastic fluid network considered here (with a deterministic cdf G) is a simplification (limiting form) of a network of infinite-server queues with batch arrivals. For relevant background on single infinite-server queues with batch arrivals, see Shandbhag [36], Holman, Chaudhry and Kashyap [18], Chapter 5 of Chaudhry and Templeton [6], Liu, Kashyap and Templeton [29], Liu and Templeton [30] and references therein. These models in turn are related to shot noise processes; see Daley [7], Kluppelberg and Mikosch [27], Rice [34], Vervaat [37] and references therein. Especially tractable are fluid analogs of a network of $M^X/M/\infty$ queues, which we call the Lévy stochastic fluid network with proportional release rates. In the fluid network the input is assumed to be an m-dimensional nondecreasing Lévy process (a nondecreasing right-continuous process with stationary independent increments). For this model, we obtain a relatively tractable expression for its steady-state distribution (Section 5). For related work when the driving process is a Brownian motion, see Jacobsen [21].

We are also able to describe the time-dependent mean fluid content of each queue in very general time-dependent stochastic fluid networks by making connections to previous results for networks of infinite-server queues with nonhomogeneous Poisson arrival processes (denoted by M_t). In remarkable generality, we show that the mean fluid contents are identical to the mean queue lengths derived in Massey and Whitt [31] (Sections 2–3).

A key structural property is the linearity (additivity). If W^1 and W^2 are content processes associated with input processes A^1 and A^2, then $W^1 + W^2$ is the content process associated with the input process $A^1 + A^2$. Thus, extensions to multiple-class networks are immediate.

Because of the tractable expressions for performance measures, the linear stochastic fluid model is very promising for applications, just like infinite-server queues. Infinite-server queues are appealing as approximations for multi-server queues, with an infinite waiting room, a finite waiting room or no waiting room. See Whitt [38], Massey and Whitt [32], Jennings, Mandelbaum, Massey and Whitt [19], Jennings and Massey [20], Grier, Massey, McKoy and Whitt [17], Duffield and Whitt [11] and references therein for a discussion of approximations and applications in the queueing context. The linear stochastic fluid networks are natural models to consider in the same setting, when individual customers or jobs are small compared to the total system size.

2. The $G_t/G_t/\infty$ Fluid Model

We start by considering a single linear stochastic fluid queue. Let $A(s,t]$ be the external input during the time interval $(s,t]$, where $-\infty < s < t < \infty$; i.e., $AB = A(B) \equiv A(\omega, B)$ is modeled

as a random measure on the real line (ω is a sample point and B is a Borel set). We assume that $P[A(s,t] < \infty] = 1$ for each finite interval $(s,t]$. The t in the subscripts of G_t in $G_t/G_t/\infty$ is to emphasize the potential time inhomogeneity.

We now specify the service mechanism. Let $G = \{G(x,t)|\ (x,t) \in \Re^2\}$ be a stochastic process with $0 \le G(\cdot,\cdot) \le 1$, such that for every fixed x, $G(x,\cdot)$ is nondecreasing and right-continuous with $G(x,x-) = 0$. We stipulate that a proportion $G(x,t)$ of any input arriving at time x departs by time t. In general, $G(x,\cdot)$ need not be proper; i.e., it is allowed that some input may never leave. Furthermore, it is possible that $G(x,x) > 0$ or even $G(x,x) = 1$, i.e. it is allowed that some, or even all, arriving input may leave as soon as it arrives. In addition, we assume that for each t, $G(\cdot,t)$ has left-continuous sample paths. As for the input process, the t in the service G_t is to point out the possibility of time inhomogeneity of the service mechanism. One example of such a process is $G(x,t) = H(t-x)$ where H is a deterministic cdf with $H(0-) = 0$. Another example is $G(x,t) = 1 - H^c(t)/H^c(x)$, where $H^c(\cdot) = 1 - H(\cdot)$ and H is a deterministic cdf. A generalization of the second example is $G(x,t) = 1 - e^{-\int_x^t R(u)du}$ where R is a nonnegative stochastic process. It is identical to the second example when H has a density h and we take $R(t) = h(t)/H^c(t)$.

Letting $G^c = 1 - G$, $\tilde{G}(x,t) = 1 - G^c(x,t)/G^c(x,x)$, $\tilde{G}^c = 1 - \tilde{G}$, $\tilde{A}(B) = \int_B \tilde{G}^c(x,x)dA(x)$ (the random measure associated with arriving input that does not instantly leave) and denoting the initial content by $W(0) = A\{0\}G^c(0,0) = \tilde{A}\{0\}$ (the remaining net value of a potential batch that arrived at time zero), we define the workload (buffer content) at time t and the output (departures) in the interval $[0,t]$ by

$$W(t) = \int_{[0,t]} G^c(x,t)dA(x) = W(0)\tilde{G}^c(0,t) + \int_{(0,t]} \tilde{G}^c(x,t)d\tilde{A}(x)\ , \quad t \ge 0\ , \tag{2.1}$$

and

$$D(t) = \int_{[0,t]} G(x,t)dA(x) = \int_{[0,\infty)} G(x,t)dA(x)\ , \quad t \ge 0\ , \tag{2.2}$$

respectively. From (2.2) and the fact that G is nondecreasing, right-continuous and bounded, it follows by bounded convergence that D is a right-continuous nondecreasing process (when converging from the right to t, let the integral in (2.2) be on any fixed finite open interval containing $[0,t]$, so that its A-measure is finite). We understand the stochastic integrals in (2.1) and (2.2) to be defined for each sample path. Since $G(\cdot,t)$ (hence, $G^c(\cdot,t)$) is left continuous, (2.1) and (2.2) are well defined as Riemann-Stieltjes integrals for each sample path; see Chapter 9 of Apostol [1],

especially p. 200. Thus, the integrals can be represented as limits of finite sums, i.e.,

$$\int_{(0,t]} G(x,t)dA(x) = \lim_{n\to\infty} \sum_{k=0}^{n-1} G\left(\frac{kt}{n}, t\right) A\left(\frac{kt}{n}, \frac{(k+1)t}{n}\right] \ . \tag{2.3}$$

Note that we have the basic conservation relation

$$W(t) = A[0,t] - D(t) \ , \quad t \geq 0 \ , \tag{2.4}$$

as can be seen by adding equations (2.1) and (2.2). Thus, W is a right-continuous bounded-variation (hence having left limits) process, which from (2.1) is also nonnegative.

When $G(x,t) = e^{-\int_x^t R(u)du}$, the process $W(t)$ in (2.1) is the unique process satisfying

$$W(t) = A[0,t] - \int_0^t R(s)W(s)ds \ . \tag{2.5}$$

That is, W is a dam process with a modulated linear release rate. See Sections 3 and 4 of Asmussen and Kella [3] for more on this case.

Under minor regularity conditions, it is easy to give expressions for mean values.

Lemma 2.1 *In the $G_t/G_t/\infty$ fluid model, assume that A and G are independent, that $G_0(s) \equiv EG(x,x+s)$ is (functionally) independent of x, and that*

$$EA[0,t] = \beta + \int_0^t \alpha(u)du \ , \quad t \geq 0 \ . \tag{2.6}$$

Then the mean values are

$$m(t) \equiv EW(t) = \beta G_0^c(t) + \int_0^t G_0^c(t-u)\alpha(u)du \tag{2.7}$$

and

$$ED(t) = \beta G_0(t) + \int_0^t G_0(t-u)\alpha(u)du \ . \tag{2.8}$$

The Laplace transforms of the means are

$$\hat{m}(s) \equiv \int_0^\infty e^{-st} m(t)dt = (\beta + \hat{\alpha}(s))\frac{(1-\hat{g}_0(s))}{s} \tag{2.9}$$

$$\int_0^\infty e^{-st} ED(t)dt = (\beta + \hat{\alpha}(s))\frac{\hat{g}_0(s)}{s} \ , \tag{2.10}$$

where

$$\hat{g}_0(s) = \int_{[0,\infty)} e^{-st} dG_0(t) \quad \textit{and} \quad \hat{\alpha}(s) = \int_0^\infty e^{-st}\alpha(t)dt \tag{2.11}$$

If, in addition, $\beta = 0$, then $EW(t)$ and $ED(t)$ have the same form as the mean queue length at time t and mean number of departures in the interval $[0,t]$, respectively, in the $M_t/GI/\infty$ queueing model starting empty at time 0 with arrival rate function $\alpha(t)$ and service-time cdf G_0.

Proof. To obtain (2.7) and (2.8), take expectations in (2.1) and (2.2), exploiting (2.6). For the Laplace transform, use the property that the transform of a convolution is the product of the transforms. For the $M_t/GI/\infty$ formulas, see Theorem 1 of Eick et al. [12]. ∎

We remark that the $M_t/GI/\infty$ mean formulas do not depend on the Poisson property; see Remark 2.3 of Massey and Whitt [31]. Lemma 2.1 implies that, under the conditions specified, previously derived formulas for the means in the $M_t/G/\infty$ queue in Eick et al. [12, 13] and Massey and Whitt [31] also apply to the means in $G_t/G_t/\infty$ fluid queues. We thus have many explicit mean formulas for the $G_t/G_t/\infty$ fluid model. Important examples are the cases in which the arrival-rate function $\alpha(t)$ is polynomial or sinusoidal.

If G_0 from Lemma 2.1 is absolutely continuous with density g_0, then the departure rate is well defined, i.e.,

$$ED(t) = \int_0^t \delta(s)ds \ , \quad t \geq 0 \ , \tag{2.12}$$

where

$$\delta(t) = \beta g_0(t) + \int_0^t g_0(t-x)\alpha(x)dx \ . \tag{2.13}$$

We note that for the case where $\alpha(t) = \alpha > 0, \ t \geq 0$ (in particular when the input is a stationary random measure, i.e., has stationary increments), then

$$EW(t) = \beta G_0^c(t) + \alpha \int_0^t G_0^c(x)dx \to \alpha m \ \text{ as } \ t \to \infty \ , \tag{2.14}$$

where m is the mean of the cdf G_0 (possibly infinite). Note that when m is finite $G_0^c(t)$ necessarily vanishes as $t \to \infty$.

Now suppose that in addition to the conditions of Lemma 2.1 $\{A(0,t] : t \geq 0\}$ has independent increments, $A\{0\}$ and G are deterministic (so that necessarily, $G(x,t) = G_0(t-x)$ for all x, t), and

$$VarA(0,t] = V(t) < \infty \ , \quad t \geq 0 \ . \tag{2.15}$$

Then

$$VarW(t) = \int_{[0,t]} G_0^c(t-x)^2 dV(x) \tag{2.16}$$

and

$$VarD(t) = \int_{[0,t]} G_0(t-x)^2 dV(x) \ , \quad t \geq 0 \ , \tag{2.17}$$

as can be seen from (2.3).

For example, if $\{A(0,t] : t \geq 0\}$ is a nonhomogeneous compound Poisson process, i.e., if

$$A(0,t] = \sum_{i=1}^{N(t)} X_i \ , \quad t \geq 0 \ , \tag{2.18}$$

where X_i are i.i.d. and $N(t)$ is a nonhomogeneous Poisson process with intensity function $\lambda(t)$, then

$$V(t) \equiv VarA(t) = E(X^2) \int_0^t \lambda(u)du \ , \quad t \geq 0 \ , \tag{2.19}$$

and

$$VarW(t) = E(X^2) \int_0^t G_0^c(t-x)\lambda(x)dx \ . \tag{2.20}$$

If in addition $\lambda(t) = \lambda t \geq 0$, then

$$VarW(t) \to m\lambda E(X^2) \ \text{ as } t \to \infty \ , \tag{2.21}$$

where again m is the mean of G_0. For the stationary input case, in the limit as $t \to \infty$, $W(t)$ then has mean $m\lambda E(X)$ and variance $m\lambda E(X^2)$. Only in the pure Poisson case ($X = 1$) is the mean equal to the variance.

Denote $A^u(B) = A(u + B)$ and $G^u(x,t) = G(x + u, t + u)$, where $u + B = \{u + b : b \in B\}$. We now state the following general result.

Theorem 2.1 *Assume that the (joint) distribution of (G^u, A^u) is independent of u. Then $A\{0\} = 0$ a.s., and if*

$$W^*(0) = \int_{(-\infty,0]} G^c(x,0)dA(x) \tag{2.22}$$

is a.s. finite, then

$$W^*(t) = \int_{(-\infty,t]} G^c(x,t)dA(x) \ , \quad t \geq 0 \ , \tag{2.23}$$

is a stationary process. Furthermore, $W(t)$ converges in distribution to $W^*(0)$ and is stochastically increasing. If $G_0(\cdot) = EG(0,\cdot)$ is a proper cdf ($G_0(t) \to 1$ as $t \to \infty$), then changing the initial condition to a nonnegative a.s. finite random variable (i.e., replacing $A(\cdot)$ in (2.1) by $A(\cdot)+\xi 1_{\{0\}}(\cdot)$ where $P[0 \leq \xi < \infty] = 1$) will not change the limiting distribution. Finally, if in addition A and G are independent, $\alpha = EA(0,1] < \infty$ and G_0 has a finite mean m, then $EA(s,t] = m(t-s)$ for $s < t$ and $EW^*(0) = \alpha m < \infty$ (hence, $W^*(0)$ is a.s. finite).

Proof. A random measure can have at most a countable number of fixed atoms (e.g., Prop. 6.3.III, p. 173 of [8]). If 0 is a fixed atom of A then by stationarity any t is such, which is a contradiction. Thus $P[A\{0\} = 0] = 1$. Note that for every u, t,

$$W^*(u+t) = \int_{(-\infty,u+t]} G^c(x,u+t)dA(x) = \int_{(-\infty,t]} G^{u,c}(x,t)dA^u(x) \ , \tag{2.24}$$

where $G^{u,c} = 1 - G^u$, which implies that for every $n \geq 1$ and $t_1 < \ldots < t_n$, $(W^*(t_1+u), \ldots, W^*(t_n+u))$ is a functional of (G^u, A^u), so that stationarity clearly follows from the conditions of the theorem.

The fact that $W(t)$ is stochastically increasing and converges in distribution to $W^*(0)$ is immediate from

$$W(t) = \int_{[0,t]} G^c(x,t)dA(x) = \int_{[-t,0]} G^{t,c}(x,0)dA^t(x) \sim \int_{[-t,0]} G^c(x,0)dA(x) \ . \tag{2.25}$$

Next, we observe that since $G(0,t)$ is bounded above by 1, $G_0(t) \to 1$ if and only if $G(0,t) \to 1$ a.s. Thus, for any a.s. finite random variable ξ, $G^c(0,t)\xi$ vanishes a.s. as $t \to \infty$, which gives the insensitivity to initial conditions. Finally, $EA(s,t] = \alpha(t-s)$ follows from the additivity and monotonicity of $EA(0,\cdot]$, and when A and G are independent $EW^*(0) = \alpha m < \infty$ follows from Lemma 2.1 and (2.14). ∎

Note that A in Theorem 2.1 is necessarily a stationary random measure. Two examples for which the time-stationarity condition of Theorem 2.1 is satisfied are

1. $G(x,t) = H(t-x)$ and H is a stochastic cdf with $H(0-) = 0$ which is independent of A and A is a stationary random measure (has stationary increments).

2. $G(x,t) = 1 - e^{\int_x^t R(s)ds}$, where the distribution of (R^u, A^u) is independent of u with $R^u(t) = R(u+t)$. In particular, R is a stationary process and A is a stationary random measure. This is seen by observing that $G^u(x,t) = 1 - e^{\int_x^t R^u(s)ds}$. See [3] for more on this and related processes.

We now further characterize the steady-state distribution when the input process is compound Poisson and $G(x,t) = G_0(t-x)$ is deterministic.

Theorem 2.2 *Let $A(0,\cdot]$ be a homogeneous compound Poisson process as in (2.18) with Poisson arrival rate λ and random jump sizes X_i and assume that $G(x,t) = G_0(t-x)$ is deterministic. Let S_n denote the n^{th} point of the Poisson process to the right of the origin. Then the stationary workload in (2.23) is distributed as the "shot noise" series*

$$W^*(0) = \sum_{n=1}^{\infty} X_n G_0^c(S_n) \ , \tag{2.26}$$

which is infinitely divisible with mean $m\lambda EX$ and variance $m\lambda EX^2$, where m is the mean of G_0. Moreover, taking $\eta(\alpha) = \lambda(1 - Ee^{-\alpha X_1}) = -\log(Ee^{-\alpha A(0,1]})$, the Laplace-Stieltjes transform of $W^*(0)$ is given by

$$Ee^{-\alpha W^*(0)} = e^{-\int_0^\infty \eta(G_0^c(x))dx} \ . \tag{2.27}$$

Proof. The representation (2.26) follows directly from (2.23) and (2.18). The property of being infinitely divisible is discussed on p. 766 of Vervaat [37]. The rest is straightforward. To show (2.27), see (5.9) below and the sentence that follows. ∎

In addition to the steady state workload at an arbitrary time, we can consider the steady state seen by an arrival (not counting the input at that arrival epoch). If the arrival process A is a Lévy process, then these two notions of steady state coincide (the LASTA property); see Remark 5 on p. 162 of Melamed and Whitt [33].

3. Extension to Networks

We start our extension to networks by considering m fluid queues in tandem (series) with external input A to the first queue, denoted by $G_t/G_{t,1}/\infty \to G_{t,2}/\infty \to \ldots \to G_{t,m}/\infty$. Let the service mechanism G_k govern the output from queue k; i.e., a proportion $G_k(x,t)$ of all input that arrives to queue k at time x departs by time t. Let $H_0(x,t) = 1_{[x,\infty)}(t)$ and $H_k(x,t) = \int_{[x,t]} G_k(y,t) H_{k-1}(x,dy)$. In particular, $H_1 = G_1$ and if $G_k(x,t) = G_{0,k}(t-x)$ (deterministic), then $H_{0,k}(\cdot) \equiv H_k(0,\cdot)$ is the k-fold convolution of $G_{0,1}, \ldots, G_{0,k}$. With the above setup, the workload and output processes for the k^{th} queue are

$$W_k(t) = \int_{[0,t]} [H_{k-1}(x,t) - H_k(x,t)] dA(x) \ , \tag{3.1}$$

and

$$D_k(t) = \int_{[0,t]} H_k(x,t) dA(x) \ , \quad t \geq 0 \ . \tag{3.2}$$

To see this, note that with $D_0 = A$ we have that $D_k(t) = \int_{[0,t]} G_k(x,t) dD_{k-1}(x)$ and $W_k(t) = \int_{[0,t]} G_k^c(x,t) dD_{k-1}(x)$ for all $1 \leq k \leq m$.

Paralleling Theorem 2.1, when the distribution of $(G_1^u, \ldots, G_m^u, A^u)$ is independent of u (so that in particular A is a stationary random measure, hence $A\{0\} = 0$ a.s.) it is easy to check that the distribution of $(H_1^u, \ldots, H_m^u, A^u)$ is independent of u, where $H_k^u(x,t) = H_k(u+x, u+t)$. In this case we can construct a stationary version of $\{(W_1(t), \ldots, W_m(t)) : t \geq 0\}$, i.e.,

$$W_k^*(t) = \int_{(-\infty,t]} [H_{k-1}(x,t) - H_k(x,t)] dA(x) \ , \quad t \geq 0 \ , \tag{3.3}$$

provided that $W_k^*(0)$ is a.s. finite for all k. Suppose that $G_k(x,t) = G_{0,k}(t-x)$ and $A\{0\}$ are deterministic and that $A(0,\cdot]$ has independent increments, with (2.15) holding. Then the workload variances are

$$VarW_k(t) = \int_{[0,t]} [H_{0,k-1}(t-x) - H_{0,k}(t-x)]^2 dV(x) \ . \tag{3.4}$$

Moreover, the covariance between workloads (possibly at different times) have the form

$$Cov(W_j(t_1), W_k(t_2)) = \int_{[0,t_1 \wedge t_2]} [(H_{0,j-1}(t_1-x) - H_{0,j}(t_1-x))(H_{0,k}(t_2-x) - H_{0,k}(t_2-x)]dV(x) \quad (3.5)$$

where $t_1 \wedge t_2 = \min(t_1, t_2)$.

We now consider general networks with deterministic routes. For any deterministic route, we can initially treat revisits to the same queue as visits to different queues by relabeling every revisited queue as a distinct queue. Then we can apply the tandem queue results above. Afterwards we can add over those instances where a queue is repeated. For a tandem model, we can recursively apply the results of Section 2. There we showed that the mean workload at one queue and mean output from that queue have the same form as the mean number of busy servers and the mean number of departures in the $M_t/G/\infty$ queue. Hence, we can proceed recursively to successive queues in the tandem model. Finally suppose that there is proportional routing; i.e., a probability matrix P such that a proportion P_{ij} of the output from node i goes next to node j. Then there are countably many deterministic routes through the entire network, each of which can be apportioned its portion of each originating flow.

We now construct the workload processes, using a minor modification of the construction on p. 220 of Massey and Whitt [31]. In general, we have a countable collection T of possible routes r. Associated with each route r is an initial queue $q_1(r)$. Associated with each input at queue $q_1(r)$, a proportion of that input $p(r)$ follows route r. Let $\bar{S}_k^r$ be a random variable with a cdf equal to the convolution of the first k cdf's on route r. Let $r^{-1}(k)$ be the set of sites on route r that are queue k.

Now assume in addition that for the j^{th} input A_j along the route

$$EA_j(s,t] = \int_s^t \alpha_j(s)ds \ , \quad t \geq 0 \ , \quad (3.6)$$

so that the intensity of input following route r is $\alpha^r(t) = p(r)\alpha_{q_1(r)}(t)$. We now show that the mean workload at queue k has the same form as in (5.4) and the proof of Theorem 5.3 of Massey and Whitt [31].

Theorem 3.1 *Under (3.6) the mean workload at queue k starting empty in the distant past is*

$$EW_k(t) = \sum_{r \in R} \sum_{i \in r^{-1}(k)} \int_{-\infty}^t (\bar{S}_{i-1}^r < t-s \leq S_i^r)\alpha^r(s)ds \ . \quad (3.7)$$

Proof. Following p. 220 of Massey and Whitt [31], we can define the workload at queue k starting empty in the distant past by the stochastic integral

$$W_k(t) = \sum_{r \in R} \sum_{i \in r^{-1}(k)} p(r) \int_{-\infty}^{t} P(\bar{S}_{i-1}^r < t - s \le \bar{S}_i^r) dA_{q_1}(r)(s) \ , \tag{3.8}$$

from which (3.7) is an elementary consequence. Since we are calculating means, the dependence among input processes A_j need not be considered. ∎

As a consequence of the reasoning above, we can treat the cases in which the time-dependent intensity function for each route is a polynomial or sinusoidal function; see Theorems 6.2 and 6.3 of Massey and Whitt [31].

Unfortunately, we cannot proceed beyond the means with the argument above. We can exploit (3.8) to characterize the vector workload process, but its full distribution will not be the same as for the vector queue-length distribution in the $M_t/G/\infty$ queueing networks.

4. Stochastic Fluid Networks with Proportional Release Rates

An alternative fluid model has state-dependent release rates. Let $f \equiv (f_1, \ldots, f_m)$ be a vector of nonnegative measurable real-valued functions on $[0, \infty)$. When queue k has content x, the output rate is $f_k(x)$. Let P be a substochastic transition matrix with $P^n \to 0$ as $n \to \infty$. Thus P has spectral radius less than 1, $I - P$ is nonsingular and $(I - P)^{-1} = \sum_{n=0}^{\infty} P^n$. We interpret P_{ij} as the proportion of output from queue i that goes next to queue j. (In the setting of Sections 3 and 4, $P_{ij} = 1$ or 0 for all i and j.) With this framework, the workload processes can be defined by

$$W_k(t) = W_k(0) + A_k(0,t] + \sum_{j=1}^{m} P_{jk} \int_0^t f_j(W_j(s)) ds - \int_0^t f_k(W_k(s)) ds \ , \tag{4.1}$$

$1 \le k \le m$, where $A \equiv (A_1(\cdot), \ldots, A_m(\cdot))$, $1 \le k \le m\}$ is the vector of inputs (random measures), assuming that the integrals are finite.

We will focus on the special case in which $f_k(x) = r_k x$, $x \ge 0$; i.e., the output rate is proportional to the current content. Our model is thus a natural generalization of a dam (single buffer) with proportional release rule; e.g., see Chapter XIII of Asmussen [2]. In this case of proportional release rate, we will show that the model is a linear stochastic fluid network with exponential processing time cdf's; i.e., in the setting of Sections 2–3, the deterministic service mechanism G_k at queue k is $G_k(x,t) = 1 - e^{-r_k(t-x)}$, $t \ge 0$. In other words, the proportion of fluid departing by time t after arrival to queue k is $G_k(0,t)$. We can think of each particle having an independent exponential holding time at each queue.

Remark 4.1 We have noted that the proportional routing with proportions r_k correspond to the exponential cdf's. If we actually wish to use non-exponential cdf's, then we can often still exploit the same structure, because if a queue has a phase-type cdf, then we can represent it equivalently as a network where each queue has an exponential cdf. To illustrate, a mixture of two exponential cdf's could be represented by routing proportions of the input to each of two queues with the component exponential cdf's.

With proportional rates, we can rewrite (4.1) in matrix notation

$$W(t) = W(0) + A(0,t] - Q' \int_0^t W(s)ds \ , \tag{4.2}$$

where $Q = D_r(I - P)$, $D_r = diag(r)$ for $r = (r_1, \ldots, r_m)$, M' denotes the transpose of M and $\int_0^t W(s)ds$ is a vector with components $\int_0^t W_k(s)ds$.

We now show that the vector workload process is the natural generalization of (2.1) with $e^{-Q'(t-x)}$ replacing $G^c(x,t)$.

Theorem 4.1 *For every input A, the unique solution to (4.2) is*

$$W(t) = e^{-Q't}W(0) + \int_{(0,t]} e^{-Q'(t-s)}dA(s) \ , \tag{4.3}$$

i.e.,

$$W_k(t) = \sum_{j=1}^{m}([e^{-Q't}]_{kj}W_j(0) + \int_{(0,t]} [e^{-Q'(t-s)}]_{kj}dA_j(s)) \ , \tag{4.4}$$

$1 \le k \le m$.

Proof. It is straightforward to see that $W(t)$ in (4.3) satisfies (4.2); calculate $Q' \int_0^t W(s)ds$, changing the order of integration. To see that (4.3) is the unique solution, let W^1 and W^2 be two candidate solutions. Then there difference $\Delta W = W^2 - W^1$ satisfies

$$\Delta W(t) = -Q' \int_0^t \Delta W(s)ds$$

with $\Delta W = 0$. This implies that ΔW is continuous and differentiable. Thus it satisfies

$$\frac{d\Delta W(t)}{dt} = -Q'\Delta W(t)$$

with $\Delta W(0) = 0$, which implies that $\Delta W(t) = 0$ for all t (a standard results about differential equations). ∎

Remark 4.2 Note that (4.2) is the continuous-time analog of the stochastic *difference* equation

$$W_{n+1} - W_n = A_{n+1} - A_n - Q'W_n \tag{4.5}$$

or

$$W_{n+1} = B_n + C_n W_n \ , \tag{4.6}$$

where $B_n = A_{n+1} - A_n$ and $C_n = I - Q'$. For background on such stochastic difference equations, where C_n as well as B_n may be random, see Vervaat [37], Brandt [5] and Chapter 7 of Glasserman and Yao [15]. These in turn are related to the discrete-time production models of Denardo and Tang [10] and Denardo and Lee [9].

We now consider the limiting distribution of $W(t)$ as $t \to \infty$.

Theorem 4.2 *Let $W^0(t)$ be $W(t)$ with $W(0) = 0$, i.e.,*

$$W^0(t) = \int_{(0,t]} e^{-Q'(t-s)} dA(s) \ , \quad t \geq 0 \ . \tag{4.7}$$

There exists $\epsilon > 0$ such that $e^{\epsilon t}||W(t) - W^0(t)|| \to 0$ as $t \to \infty$. Consequently, if $W(t)$ converges to a proper limit as $t \to \infty$ for some initial condition, then it converges to the same limit for any a.s. finite initial condition.

Proof. First recall that for a square matrix M the matrix exponential is

$$e^{Mt} = \sum_{k=0}^{\infty} \frac{M^k t^k}{k!} \ . \tag{4.8}$$

It is well known and easy to check (applying a Jordan decomposition of M) that the entries of e^{Mt} are finite linear combinations (possibly with complex coefficients) of functions of the form $e^{\lambda t} t^k$, where λ is a (possibly complex) eigenvalue of the matrix M and k is less than the dimension of the matrix. Hence, if the real parts of all eigenvalues of M are strictly negative, then there exists a positive ϵ such that $e^{\varepsilon t} e^{Mt} \to 0$ as $t \to \infty$. The proof is completed by noting that all eigenvalues of $-Q'$ have negative real parts, and applying Theorem 4.1. To substantiate the required property of $-Q'$, note that $P - I$ is the (sub) generator of a Markov process, which is terminating when the chain corresponding to P is, but we have assumed that P is substochastic with $P^n \to 0$. The Markov process with subgenerator $D_r(P - I)$ is also terminating, since it is just a time transformation of the other Markov process. Hence its state probability matrix $e^{D_r(P-I)t} = e^{-Qt}$ converges to zero, which in turn implies the eigenvalue property.

As in Section 2, it is now interesting to consider some additional assumptions which will insure that $W^0(t)$ converges in distribution to some proper limit or, alternatively, that there exists some choice of initial condition which will make the process stationary. In the latter case the stationary

version will trivially have a limiting distribution and thus with any a.s. finite initial condition, we will get the same limiting distribution. With this in mind, we have the following result.

Theorem 4.3 *Assume that A is a stationary random measure (on $\Re^m$), and that $EA_i(0,1] = \alpha_i < \infty$ for all $i = 1, ..., m$. Then, for every $t \geq 0$, $W^0(t)$ in (4.7) and $\int_{(-t,0]} e^{Q's} dA(s)$ are identically distributed, so that $\{W^0(t) | t \geq 0\}$ is stochastically increasing. Furthermore, let*

$$W^*(0) \equiv \int_{(-\infty,0]} e^{Q's} dA(s) \tag{4.9}$$

and let W^* be the solution of

$$W^*(t) = W^*(0) + A(0,t] - Q' \int_0^t W^*(s) ds \ . \tag{4.10}$$

Then the mean vector of $W^*(0)$ is finite and is given by

$$EW^*(0) = (Q')^{-1} \alpha \ , \tag{4.11}$$

where $\alpha = (\alpha_i)$. Moreover, W^* is a stationary process and, for any a.s. finite initial condition, $W(t)$ converges in distribution to $W^*(0)$ as $t \to \infty$.

Proof. The fact that $W^0(t)$ and $\int_{(-t,0]} e^{-Q's} dA(s)$ are identically distributed holds since

$$W^0(t) = \int_{(0,t]} e^{-Q'(t-s)} dA(s) = \int_{(-t,0]} e^{Q's} dA^t(s) \sim \int_{(-t,0]} e^{Q's} dA(s) \tag{4.12}$$

where $A^t(B) = A(t + B)$. Stochastic monotonicity is now implied by the nonnegativity of the matrix $e^{Q's}$ for $s \leq 0$. To proceed, we first note that since $EA_i(0, \cdot]$ is additive and monotone, then $EA_i(t) = \alpha_i t$. In order to show that $EW^*(0) < \infty$, by the proof of Theorem 4.2, it suffices to show that for any $\varepsilon > 0$ and any $i = 1, \ldots, n$ we have that $E \int_{(-\infty,0]} e^{\varepsilon s} dA_i(s) < \infty$. Integration by parts gives

$$\begin{array}{rcl} E \int_{(t,0]} e^{\varepsilon s} dA_i(s) & = & E \left(-e^{\varepsilon t} A_i(0,t] + \varepsilon \int_0^t e^{\varepsilon s} A_i(0,s] ds \right) \\ & = & -e^{\varepsilon t} \alpha_i + \varepsilon \int_0^t e^{\varepsilon s} \alpha_i s ds = \alpha_i \int_t^0 e^{\varepsilon s} ds \ . \end{array} \tag{4.13}$$

By monotone convergence, this implies that $E \int_{(-\infty,0]} e^{\varepsilon s} dA_i(s) = \alpha_i / \varepsilon < \infty$. To show that W^* is a stationary process, we apply (4.3) to give

$$W^*(t) = \int_{(-\infty,t]} e^{-Q'(t-s)} dA(s) = \int_{(-\infty,0]} e^{Q's} dA^t(s) \ , \tag{4.14}$$

where the stationarity of the right hand side is given by the stationarity of A. The fact that convergence in distribution holds for any initial condition follows from Theorem 6.2. Finally, (4.11) can be shown either directly from (4.9) or a bit more simply from (4.10). In particular, in the latter, taking expectations, applying stationarity and subtracting $EW^*(0) = EW^*(t)$ from both sides, one obtains $0 = \alpha t - Q' EW^*(0) t$, which gives (4.11). ∎

5. Lévy Stochastic Fluid Networks with Proportional Release Rates

As in the last section, we consider the linear stochastic fluid network governed by proportional release rate vector r and proportional routing matrix P, but now we assume in addition that the right-continuous nondecreasing process $A(0,\cdot]$ is a Lévy process. That is, it has stationary independent increments. Let $\eta(\beta) \equiv -\log(Ee^{-\beta' A(0,1]}$ be its exponent $(0 \leq \beta \in \Re^m)$.

The prototype nondecreasing Lévy process is a compound Poisson process with a possible vector valued drift $c \geq 0$, i.e., let $\{N(t) : t \geq 0\}$ be a Poisson process with rate λ and let X_i, $i \geq 1$, be i.i.d. nonnegative random vectors with Laplace transform

$$\psi(\beta) \equiv Ee^{-\beta' X_1} \equiv Ee^{-\sum_{j=1}^m \beta_j X_{1j}} \ . \tag{5.1}$$

Then

$$A(0,t] = ct + \sum_{i=1}^{N(t)} X_i \tag{5.2}$$

and for this special case the exponent is

$$\eta(\beta) = \beta' c + \lambda(1 - \psi(\beta)) \ . \tag{5.3}$$

Any nondecreasing Lévy process is actually a limit of such compound Poisson processes with nonnegative jumps and nonnegative drifts.

The Lévy structure enables us to obtain very attractive explicit results for the steady-state distribution.

Theorem 5.1 *If, in addition to the conditions of Theorem 4.3, $A(0,\cdot]$ is a multivariate Lévy process with exponent $\eta(\beta) = -\log Ee^{-\beta' A(0,1]}$, which is independent of $W(0)$, then $W(t)$ is distributed as*

$$e^{-Q't} W(0) + \int_{(0,t]} e^{-Q's} dA(s) \ , \tag{5.4}$$

Thus

$$\tilde{W}(\beta, t) \equiv Ee^{-\beta' W(t)} = \tilde{W}(e^{-Qt}\beta, 0) e^{-\int_0^t \eta(e^{-Qs}\beta) ds} \tag{5.5}$$

$$\to \tilde{W}^*(\beta) \equiv Ee^{-\beta' W^*(0)} = e^{-\int_0^\infty \eta(e^{-Qs}\beta) ds} \tag{5.6}$$

as $t \to \infty$. In particular, $W^*(0)$ has an infinitely divisible distribution and if β_λ is a right eigenvector of Q with respect to some eigenvalue λ, then

$$\tilde{W}^*(\beta_\lambda) = e^{-\int_0^1 \frac{\eta(x\beta_\lambda)}{\lambda x} dx} \tag{5.7}$$

Moreover, $\tilde{W}^*(\beta)$ solves the following differential equation

$$\beta' Q' \nabla \log \tilde{W}^*(\beta) = -\eta(\beta) \ . \tag{5.8}$$

Proof. The fact that (5.4) holds follows from (4.3) via an obvious change of variable ($s := t - s$) recalling the fact that, since $A(0, \ldots]$ is Lévy, $\{A(0,s]|0 \le s \le t\}$ and $\{A(t-s,t]|0 \le s \le t\}$ are identically distributed processes. Next, we observe that for any function $h : \Re \to \Re^n$ which is a.s. continuous (w.r.t. Lebesgue measure) and bounded on finite intervals,

$$E e^{-\int_0^t h(s)' dA(s)} = e^{-\int_0^t \eta(h(s))ds} \ . \tag{5.9}$$

From the stationary independent increments property, it is easy to show that it holds for (vector valued) step functions and thus also for functions which are a.s. continuous and bounded (i.e., h and any continuous function of h is Riemann integrable) on finite intervals. In particular this holds for $h(s) = e^{-Qs}\beta$, which together with (5.4) gives (5.5). Infinite divisibility of $W^*(0)$ follows from (4.9) and the fact that, for every n, the sum of n i.i.d. Lévy processes each with exponent η/n is a Lévy process with exponent η. Now, (5.7) follows from (5.5) by the fact that $e^{-Qs}\beta_\lambda = e^{-\lambda s}\beta_\lambda$ and the change of variables $x := e^{-\lambda s}$. As for (5.8) we first observe that

$$\nabla \log \tilde{W}^*(\beta) = -\int_0^\infty e^{-Q's} \nabla \eta(e^{-Qs}\beta) ds \ . \tag{5.10}$$

Since $e^{-Qt} \to 0$ as $t \to \infty$,

$$-\int_0^\infty \beta' Q' e^{-Q's} \nabla \eta(e^{-Qs}\beta) ds = \eta(e^{-Qs}\beta)\Big|_0^\infty = -\eta(\beta) \ , \tag{5.11}$$

which implies (5.8). ∎

We now give an expression for the covariance matrix under extra regularity conditions. See Theorem 2.12 of Jacobsen [21] for a similar result for the case $A(0,t]$ is a Brownian motion.

Theorem 5.2 *If, in addition to the conditions of Theorem 5.1, $A(0,1]$ has finite second moments with covariance matrix $\Sigma = -\nabla^2 \eta(0) = (\sigma_{ij})$, then the covariance matrix $\Sigma^* = (\sigma^*_{ij})$ of $W^*(0)$ is given by*

$$\Sigma^* = \int_0^\infty e^{-Q's} \Sigma e^{-Qs} ds \tag{5.12}$$

and is the unique solution of

$$Q'\Sigma^* + \Sigma^* Q = \Sigma \ . \tag{5.13}$$

In particular, for the two dimensional case,

$$\begin{aligned}
\sigma_{12}^* &= \frac{\sigma_{11}p_{12} + \sigma_{22}p_{21} + 2\sigma_{12}}{2(r_1 + r_2)(1 - p_{12}p_{21})} \\
\sigma_{11}^* &= \frac{\sigma_{11}}{2r_1} + \frac{r_2 p_{21}}{r_1}\sigma_{12}^* \\
\sigma_{22}^* &= \frac{\sigma_{22}}{2r_2} + \frac{r_1 p_{12}}{r_2}\sigma_{12}^*
\end{aligned} \tag{5.14}$$

Proof. From (5.5) we have that

$$\begin{aligned}
\nabla^2 \log \tilde{W}^*(\beta) &= \frac{\nabla^2 \tilde{W}^*(\beta)}{\tilde{W}^*(\beta)} - \frac{\nabla \tilde{W}^*(\beta) \nabla \tilde{W}^*(\beta)'}{\tilde{W}^*(\beta)^2} \\
&= -\int_0^\infty e^{-Q's} \nabla^2 \eta(e^{-Qs}\beta) e^{-Qs} ds
\end{aligned} \tag{5.15}$$

so that (5.12) is obtained by letting $\beta \to 0$. To show (5.13) we observe that

$$-\frac{d}{dt} e^{-Q't} \Sigma e^{-Qt} = Q' e^{-Q't} \Sigma e^{-Qt} + e^{-Q't} \Sigma e^{-Qt} Q \tag{5.16}$$

and the result is obtained by integrating both sides and recalling that $e^{-Qt} \to 0$ as $t \to \infty$. To show that the solution of (5.13) is unique, we first observe that the difference between any two solutions is some symmetric matrix S which satisfies $Q'S + SQ = 0$. This implies that $(Q')^n S = S(-Q)^n$ and, by multiplying by $t^n/n!$ and summing, that $e^{Q't} S = S e^{-Qt}$. Thus, $S = e^{-Q't} S e^{-Qt}$ which vanishes as $t \to \infty$. Therefore, $S = 0$ and the solution is unique. Finally, (5.14) is obtained by explicitly solving the linear system of three equation in three unknowns dictated by (5.13). We note that σ_{12}^* on the right hand side of the bottom two equations is not an error. ∎

6. Concluding Remarks

We conclude with some remarks about applications to control and heavy-traffic limit theorems.

Centralized and Distributed Control

The linear stochastic fluid model provides a way to combine centralized and distributed control. The central controller might choose the proportions $P \equiv (P_{ij})$ and/or $r \equiv (r_j)$. Then a local controller at queue k might take actions throughout time to achieve processing at rate $r_k x$ whenever his workload is x. Moreover, he can route his output according to the specified proportions P_{kj}, $1 \le j \le m$.

Given projected input according to the input A with intensity vector α, the controller might aim to minimize his expected work in process, which is given by $Q'^{-1}\alpha$ in (4.11). If P is given, then the mean total workload is

$$EW = \sum_{i=1}^{m} \lambda_i / r_i \ , \tag{6.1}$$

where the net input rates λ_i are obtained by solving the traffic rate equations

$$\lambda_i = \alpha_i + \sum_{j=1}^{m} \alpha_j P_{ij} \ , \quad 1 \le i \le m \ . \tag{6.2}$$

The central controller might choose to minimize EW over possible rate vectors r subject to a constraint on costs associated with the rates

$$\sum_{i=1}^{m} r_i c_i \le C \ . \tag{6.3}$$

Clearly the mean is reduced by increasing r_i, so that it suffices to consider an equality constraint in (6.3). The elementary solution to this problem is

$$r_i = \frac{C\sqrt{\lambda_i / c_i}}{\sum_{j=1}^{m} \sqrt{\lambda_j / c_j}} \ , \quad 1 \le i \le m \ . \tag{6.4}$$

Heavy-Traffic Limit Theorems

Paralleling the heavy-traffic limits for infinite-server queues in Glynn and Whitt [16], we can obtain heavy-traffic limits for linear stochastic fluid networks. Indeed, we can apply the previous theorems because the previous theorems did not depend on the input processes being integer valued. Those theorems depend on the service-time distribution being a finite mixture of atoms. However, here, if the processing-time cdf G is associated with finitely many atoms, i.e., if

$$G(t) = \sum_{i=1}^{k} p_i 1_{[0,x_i]}(t) \ , \quad t \ge 0 \ , \tag{6.5}$$

then we understand a deterministic proportion p_i of the arrivals to require x_i units of processing time. Hence, the workload at time t is

$$W(t) = \sum_{i=1}^{k} p_i A(t - x_i, t] \ , \tag{6.6}$$

i.e., in the setting of Glynn and Whitt [16], $N^i = p_i N$. We can thus apply the theorems there to get limits as $n \to \infty$ for normalized processes such as

$$n^{\gamma}(n^{-1}W(nt) - m(t)) \ , \quad t \ge 0 \ , \tag{6.7}$$

The arrival rate grows by assuming that

$$n^{\gamma}(n^{-1}A(0,nt] - \lambda t) \Rightarrow Z(t) \text{ as } n \to \infty \ , \tag{6.8}$$

as in (2.3) of Glynn and Whitt [16]. Under regularity conditions, these limits support Gaussian approximations with the exact means and covariances when the input rate is relatively high.

References

[1] Apostol, T. M. (1957) *Mathematical Analysis*, Addison-Wesley, Reading, MA.

[2] Asmussen, S. (1987) *Applied Probability and Queues*, Wiley, New York.

[3] Asmussen, S. and Kella, O. (1996) Rate modulation in dams and ruin problems. *J. Appl. Prob.* 33, 523-535.

[4] Berman, A. and Plemmons, R. J. (1979) *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York.

[5] Brandt, A. (1986) The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Adv. Appl. Prob.* 18, 211–220.

[6] Chaudhry, M. L. and Templeton, J. G. C. (1983) *A First Course in Bulk Queues*, Wiley, New York.

[7] Daley, D. J. (1971) The definition of a multi-dimensional generalization of shot noise. *J. Appl. Prob.* 8, 128–135.

[8] Daley, D. J. and Vere-Jones, D. (1988) *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York.

[9] Denardo, E. V. and Lee, T. Y. S. (1966) Managing uncertainty in a serial production line. *Opns. Res.* 44, 382–392.

[10] Denardo, E. V. and Tang, C. S. (1992) Linear control of a Markov production system. *Opns. Res.* 40, 259–278.

[11] Duffield, N. G. and Whitt, W. (1997) Control and recovery from rare congestion events in a large multi-server system. *Queueing Systems*, 26, 69–104.

[12] Eick, S. G., Massey, W. A. and Whitt, W. (1993a) The physics of the $M_t/G/\infty$ queue. *Opns. Res.* 41, 731–742.

[13] Eick, S. G., Massey, W. A. and Whitt, W. (1993b) $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* 39, 241–252.

[14] Feller, W. (1971) *An Introduction to Probability Theory and its Applications*, vol. II, second edition, Wiley, New York.

[15] Glasserman, P. and Yao, D. D. (1994) *Monotone Structure in Discrete-Event Systems*, Wiley, New York.

[16] Glynn, P. W. and Whitt, W. (1991) A new view of the heavy-traffic limit theorem for infinite-server queues. *Adv. Appl. Prob.* 23, 188–209.

[17] Grier, N., Massey, W. A., McKoy, T. and Whitt, W. (1997) The time-dependent Erlang loss model with retrials. *Telecommunication Systems*, 7, 253–265.

[18] Holman, D. F., Chaudhry, M. L. and Kashyap, B. R. K. (1982) On the number in the system $GI^X/M/\infty$. *Sankhyā* A44, 294–297.

[19] Jennings, O. B., Mandelbaum, A., Massey, W. A. and Whitt, W. (1996) Server staffing to meet time-varying demand. *Management Sci.* 42, 1383–1394.

[20] Jennings, O. B. and Massey, W. A. (1997) A modified-offered-load approximation for time-dependent circuit-switched networks. *Telecommunication Systems*, 7, 229–251.

[21] Jacobsen, M. (1993) A brief account of the theory of homogeneous Gaussian diffusions in finite dimensions. *Frontiers in Pure and Applied Probability, Proceedings of the Third Finnish-Soviet Symposium on Probability Theory and Mathematical Statistics*, H. Niemi, G. Hognas, A. N. Shiryaev, A. V. Melinkow (eds.), VSP Utrecht, TVP Moscow, 86–94.

[22] Kella, O. (1993) Parallel and tandem fluid networks with dependent Lévy inputs. *Ann. Appl. Prob.* 3, 682–695.

[23] Kella, O. (1996) Stability and non-product form of stochastic fluid networks with Lévy inputs. *Ann. Appl. Prob.* 6, 186-199.

[24] Kella, O. (1997) Stochastic storage networks: stationarity and the feedforward case. *J. Appl. Prob.* 34.

[25] Kella, O. and Whitt, W. (1992) A tandem fluid network with Lévy input. In *Queues and Related Models*, eds. I. Basawa and U. Bhat, Oxford University Press, Oxford, 112–128.

[26] Kella, O. and Whitt, W. (1996) Stability and structural properties of stochastic storage networks. *J. Appl. Prob.* 33, 1169–1180.

[27] Klüppelberg, C. and Mikosch, T. (1995) Explosive Poisson shot noise processes with application to risk reserves. *Bernoulli*, 1, 125–147.

[28] Kurtz, T. (1997) Limit theorems for workload input processes. In *Stochastic Networks*, eds. F. P. Kelly, S. Zachary and I. Ziedins, Oxford Publications, Oxford, 119–139.

[29] Liu, L., Kashyap, B. R. K. and Templeton, J. G. C. (1990) On the $GI^X/G/\infty$ system. *J. Appl. Prob.* 27, 671–683.

[30] Liu, L. and Templeton, J. G. C. (1991) The $GR^{X_n}/G_n/\infty$ system: system size. *Queueing Systems* 8, 323–356.

[31] Massey, W. A. and Whitt, W. (1993) Network of Infinite server queues with nonstationary Poisson input. *Queueing Systems* 13, 183–250.

[32] Massey, W. A. and Whitt, W. (1994) An analysis of the modified offered-load approximation for the nonstationary Erlang loss model. *Ann. Appl. Prob.* 4, 1145–1160.

[33] Melamed, B. and Whitt, W. (1990) On arrivals that see time averages. *Opns. Res.* 38, 156–172.

[34] Rice, J. (1977) On generalized shot noise. *Adv. Appl. Prob.* 9, 553–565.

[35] Ross, S. M. (1983) *Stochastic Processes*. Wiley, New York.

[36] Shandbhag, D.N. (1966) On the infinite-server queue with batch arrivals. *J. Appl. Prob.* 3, 274–279.

[37] Vervaat, W. (1979) On a stochastic difference equation and a representation of nonnegative infinitely divisible random variables. *Adv. Appl. Prob.* 11, 750–783.

[38] Whitt, W. (1990) Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems* 6, 335–352.