Queueing Systems 51, 77–87, 2005
© 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.

Proof of Monotone Loss Rate of Fluid Priority-Queue with Finite Buffer

STEPHEN L. SPITLER sspitler@usc.edu
Department of Electrical Engineering, University of Southern California, 3740 McClintock Avenue, Los Angeles, CA 90089-2565, USA

DANIEL C. LEE dchlee@sfu.ca
School of Engineering Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada

Received 20 September 2002; Revised 4 February 2005

Abstract. This paper studies a fluid queueing system that has a single server, a single finite buffer, and which applies a strict priority discipline to multiple arriving streams of different classes. The arriving streams are modeled by statistically independent, identically distributed random processes. A proof is presented for the highly intuitive result that, in such a queueing system, a higher priority class stream has a lower average fluid loss rate than a lower priority class stream. The proof exploits the fact that for a work-conserving queue, the fluid loss rate for a given class is invariant of what queueing discipline is applied to all arriving fluid of this particular class.

Keywords: priority queueing, monotone loss rate, finite buffer, fluid queueing system

AMS subject classification: 60K25, 68M20

1. Introduction

We consider a single-server, single-finite-buffer, fluid queueing system that applies a strict-priority queueing discipline to multiple arriving streams of different classes. Due to the finite buffer, certain entry of fluid into the queueing system can result in fluid loss due to buffer overflow. If we assume that the stochastic arrival processes associated with the fluid streams are independent and statistically indistinguishable, then it is intuitively compelling that a stream with a higher priority has a lower average loss rate than a stream with a lower priority. However, we have not seen such a proof, and we provide one in this paper.

Fluid queueing systems have been used in approximations of packet switching networks; e.g., to study congestion control [22], call admission control [1], bounds for per-flow service [15], service provisioning schemes [4,16,17], wireless transmission scheduling [24], the relative impact of queueing delays as link bandwidths become large [12], and network stability [2,11]. Markov modulated fluid flow sources have often been assumed to compute or bound queueing occupancy distributions, e.g., [5,8,13,23,25,26].

Gaussian fluid flow sources have also been used for this purpose [6,7]. Priority fluid queueing systems that have been studied include systems with separate (infinite) buffers for each class of arriving streams, e.g., [5,8,13,18,26], and systems having a single buffer with space priorities, i.e., with separate drop thresholds for each class, e.g., [9,10]. References [19,20] address queueing systems which are like that considered here in that a single buffer receives input of different priority classes, with higher priority fluid given strict priority over lower priority fluid in receiving service and buffer space. Reference [19] studies the fraction of constant-bit-rate traffic that fails to meet a delay requirement due to interference from higher priority traffic that has throughput and burstiness constraints. Reference [20] considers first passage times associated with the buffer emptying.

Section 2 elaborates on the priority-queue model that we study in this paper. Section 3 proves our result for this system, that average loss rates are monotonically non-increasing with priority. Section 4 concludes the paper.

2. System model

The fluid queueing system under consideration has finite service rate capacity μ, finite buffer size B, and is work-conserving. It receives multiple fluid streams, each of a different class. The instantaneous arrival rate of class c fluid is denoted by random process $a_c(t)$, and we allow for the possibility of bulk arrivals, i.e., $a_c(t)$ may contain impulses. The random arrival processes of all classes are assumed to be statistically indistinguishable and mutually independent; i.e., random processes $a_c(t)$, $c = 1, 2, \ldots$, are statistically independent and identically distributed. We represent the buffer occupancy of class c fluid by $q_c(t)$. At times when there is a positive work backlog ($\Sigma_c\, q_c(t) > 0$), fluid is drained from the queue at service rate μ. When the buffer is empty ($\Sigma_c q_c(t) = 0$), fluid is served at the rate, min $[\mu, \Sigma_c a_c(t)]$.

A strict priority queueing discipline is applied to the arriving fluid classes. In general, we assign positive integer labels to the classes such that priority decreases with the increase of the label value. Then the priority discipline means that, for any $j \in \{2, 3, \ldots\}$, the server cannot serve class j fluid whenever the demands of any higher priority class, labeled $c \in \{1, 2, \ldots, j - 1\}$, are not fully satisfied. These demands are due to both the higher priority fluid backlogged in the buffer and the currently arriving higher priority fluid. If there is higher priority backlogged fluid ($q_c(t) > 0$ for some $c < j$), then $\mu_j(t)$, the available service rate for class j fluid, is zero. Otherwise ($q_c(t) = 0$ for all $c < j$), $\mu_j(t)$ is that part of service rate capacity μ that is not serving fluid of the higher priority classes. Specifically, we have that

$$\mu_j(t) = \begin{cases} \mu, & j = 1 \\ \left(\prod_{c<j} 1_{\{q_c(t)=0\}}\right)\left[\mu - \sum_{c<j} a_c(t)\right]^+, & j \geq 2. \end{cases} \tag{1}$$

The strict priority applies not only to the service discipline but also to buffer usage. Letting $b_j(t)$ represent the available buffer space for class j fluid at time t, $b_j(t)$ is that part of the buffer that is not occupied by fluid of higher-priority classes; i.e.,

$$b_j(t) = \begin{cases} B, & j = 1 \\ B - \sum_{c<j} q_c(t), & j \geq 2. \end{cases} \tag{2}$$

Note that by (1) and (2), for any $j \in \{1, 2, \dots\}$ and $c \in \{j+1, j+2, \dots\}$, $\mu_j(t)$ and $b_j(t)$ are never impacted by the presence of class c fluid (i.e., $a_c(t) > 0$ or $q_c(t) > 0$ for some $c \in \{j+1, j+2, \dots\}$). The queueing dynamics of class j fluid take place exactly as if the fluid of lower-priority classes $j+1, j+2, \dots$ were not present.

3. Proof of monotonicity result

To prove that average loss rates in a strict priority queueing system are monotonically non-increasing with priority, we consider the following scenario with two such queueing systems.

3.1. Comparison of two strict priority queueing systems

Suppose that there are two fluid queueing systems, each as described in Section 2, that have the same value of μ and the same value of B. Also suppose that there are fluid streams of classes 1, 2, ..., j, where $j \geq 2$, with independent, statistically indistinguishable random arrival processes. With $a_c(t)$ denoting the arrival rate of class c fluid, $c \in \{1, 2, \dots, j\}$, at time t, we let ω_c represent a given sample path or realization drawn from the sample space for $a_c(t)$. We refer to $a_c(t)$, given realization ω_c, as $a_c(t, \omega_c)$. Also, we use $\underline{\omega}$ to represent the vector of arrival process realizations, $(\omega_1, \omega_2, \dots, \omega_j)$. One queueing system, which we call system $\langle 1 \rangle$, receives all j classes of fluid streams, i.e., classes 1, 2, ..., j. For some $i \in \{1, 2, \dots, j-1\}$, the second queueing system, system $\langle 2 \rangle$, receives i classes, i.e., classes 1, 2, ..., $i-1$, and class j. Most of our attention will be focused on the class that has lowest priority in both systems, class j.

Note that there are never more available resources for class j in system $\langle 1 \rangle$ than in system $\langle 2 \rangle$. Specifically, let $\mu_c^{\langle n \rangle}(t, \underline{\omega})$ and $b_c^{\langle n \rangle}(t, \underline{\omega})$ denote a priority queue's available service rate and available buffer space, respectively, for class c fluid, $c \in \{1, 2, \dots, j\}$, in system $\langle n \rangle$, $n \in \{1, 2\}$, at time t, given $\underline{\omega}$. Suppose initial conditions with a starting time, $t = 0$, such that, for any $\underline{\omega}$, $a_c(t, \omega_c) = 0$ for all $t < 0$ for any class, c, and that the buffers in systems $\langle 1 \rangle$ and $\langle 2 \rangle$ are both empty for all $t < 0$. (Here, we allow for bulk arrivals at time 0.) In this case, due to the strict priority discipline employed in each system, the dynamics of classes $1, 2, \dots, i-1$ are identical in systems $\langle 1 \rangle$ and $\langle 2 \rangle$; e.g., for any $\underline{\omega}$ and t, we have

$$\mu_c^{\langle 1 \rangle}(t, \underline{\omega}) = \mu_c^{\langle 2 \rangle}(t, \underline{\omega}), \quad c \in \{1, 2, \dots, i-1\}, \tag{3}$$

$$b_c^{\langle 1\rangle}(t,\underline{\omega}) = b_c^{\langle 2\rangle}(t,\underline{\omega}), \quad c \in \{1,2,\ldots,i-1\}, \tag{4}$$
$$q_c^{\langle 1\rangle}(t,\underline{\omega}) = q_c^{\langle 2\rangle}(t,\underline{\omega}), \quad c \in \{1,2,\ldots,i-1\}, \tag{5}$$

where $q_c^{\langle n\rangle}(t,\underline{\omega})$ is the buffer occupancy of class c fluid in system $\langle n\rangle$, $n \in \{1, 2\}$, at time t, for realization $\underline{\omega}$. Because there are also streams of classes $i, i+1, \ldots, j-1$ arriving at system $\langle 1\rangle$, there is always as much or more fluid with priority over class j in system $\langle 1\rangle$ than in system $\langle 2\rangle$. Therefore, the available service rates for class j in both systems are related as

$$\begin{aligned}
\mu_j^{\langle 1\rangle}(t,\underline{\omega}) &= \left(\prod_{c=1}^{j-1} 1_{\{q_c^{\langle 1\rangle}(t,\underline{\omega})=0\}}\right)\left[\mu - \sum_{c=1}^{j-1} a_c(t,\underline{\omega})\right]^+ && \text{(by (1))}\\
&\le \left(\prod_{c=1}^{i-1} 1_{\{q_c^{\langle 1\rangle}(t,\underline{\omega})=0\}}\right)\left[\mu - \sum_{c=1}^{i-1} a_c(t,\underline{\omega})\right]^+ \\
&= \left(\prod_{c=1}^{i-1} 1_{\{q_c^{\langle 2\rangle}(t,\underline{\omega})=0\}}\right)\left[\mu - \sum_{c=1}^{i-1} a_c(t,\underline{\omega})\right]^+ && \text{(by (5))}\\
&= \mu_j^{\langle 2\rangle}(t,\underline{\omega}) && \text{(by (1))}.
\end{aligned} \tag{6}$$

Similarly, by (2) and (5), the available buffer spaces for class j in both systems are such that

$$b_j^{\langle 1\rangle}(t,\underline{\omega}) \le b_j^{\langle 2\rangle}(t,\underline{\omega}). \tag{7}$$

3.2. Fluid loss over an interval

To further facilitate proof of the monotonicity result, for each class c, $c \in \{1,2,\ldots,j\}$, and for each arrival process realization ω_c, we represent each element of class c fluid uniquely by a real number, indicating the order of arrival. More specifically, we define, for each arrival process realization ω_c, a one-to-one correspondence between class c fluid arriving during time interval $[0,\infty)$ and the real half-line, $[0,\infty)$. This correspondence is ordered such that the fluid element represented by number f_1 arrives no later than that represented by f_2 if $f_1 < f_2$. In our scenario with two systems receiving the class j fluid, for each $f \in [0,\infty)$, both systems $\langle 1\rangle$ and $\langle 2\rangle$ receive the class j fluid element indexed by f, which we refer to as element f or simply f. In this representation, we define the amount of fluid represented by a set of real numbers, S, to be the volume of set S. For example, the fluid represented by the interval of real numbers, $[a,b)$, has amount $b-a$. (Here, we use intuitive notion of "volume" to avoid excessive mathematical rigor. For a higher level of mathematical rigor, "volume" can be defined, for example, as the Lebesgue measure [21] or other measures with properties satisfying the intuitive notion of fluid amount; e.g., the measure of an interval is the length of the interval, etc. With different mathematical definitions for "volume", different mathematical assumptions will have to be added for mathematical rigor.)

Letting $\underline{\omega}$ represent a given arrival process realization vector, $(\omega_1, \omega_2, \ldots, \omega_j)$, we define

$$A_j(t, \underline{\omega}) := \{f \mid \text{class } j \text{ element } f \text{ arrives during } [0, t], \text{ given realization } \underline{\omega}\}, \tag{8}$$

$$AL_j^{\langle n\rangle}(t, \underline{\omega}) := \{f \mid f \in A_j(t, \underline{\omega}) \text{ is lost (dropped from the buffer) in system } \langle n\rangle, \text{ given } \underline{\omega}\}, \quad \text{and} \tag{9}$$

$$AD_j^{\langle n\rangle}(t, \underline{\omega}) := \{f \mid f \in A_j(t, \underline{\omega}) \text{ is served in system } \langle n\rangle,\ \text{given } \underline{\omega}\}. \tag{10}$$

Note that $A_j(t, \underline{\omega})$ is partitioned into $AL_j^{\langle n\rangle}(t, \underline{\omega})$ and $AD_j^{\langle n\rangle}(t, \underline{\omega})$ in each system $\langle n\rangle$. Let $m(S)$ denote the volume of set S. Then, it immediately follows that

$$m\{A_j(t, \underline{\omega})\} = m\{AD_j^{\langle n\rangle}(t, \underline{\omega})\} + m\{AL_j^{\langle n\rangle}(t, \underline{\omega})\} \tag{11}$$

The next section is to prove the following lemma regarding the volumes of $AL_j^{\langle 1\rangle}(t, \underline{\omega})$ and $AL_j^{\langle 2\rangle}(t, \underline{\omega})$.

Lemma 1. For any $t \geq 0$, $j \in \{2, 3, \ldots\}$, and $\underline{\omega} = (\omega_1, \omega_2, \ldots, \omega_j)$,

$$m\{AL_j^{\langle 2\rangle}(t, \underline{\omega})\} \leq m\{AL_j^{\langle 1\rangle}(t, \underline{\omega})\}, \tag{12}$$

$$m\{AD_j^{\langle 1\rangle}(t, \underline{\omega})\} \leq m\{AD_j^{\langle 2\rangle}(t, \underline{\omega})\}. \tag{13}$$

Note that (12) and (13) are equivalent, which is evident from (11).

3.3. *Proof of Lemma 1*

To facilitate proof of Lemma 1, we define the following additional sets of class j fluid elements for a given arrival process realization vector, $\underline{\omega} = (\omega_1, \omega_2, \ldots, \omega_j)$:

$$L_j^{\langle n\rangle}(t, \underline{\omega}) := \{f \mid f \in AL_j^{\langle n\rangle}(t, \underline{\omega}) \text{ is lost in system } \langle n\rangle \text{ during } [0, t],\ \text{given } \underline{\omega}\}, \tag{14}$$

$$D_j^{\langle n\rangle}(t, \underline{\omega}) := \{f \mid f \in AD_j^{\langle n\rangle}(t, \underline{\omega}) \text{ is served in system } \langle n\rangle \text{ during } [0, t],\ \text{given } \underline{\omega}\}, \quad \text{and} \tag{15}$$

$$Q_j^{\langle n\rangle}(t, \underline{\omega}) := \{f \mid f \in A_j(t, \underline{\omega}) \text{ is in system } \langle n\rangle\text{'s buffer at time } t,\ \text{given } \underline{\omega}\}. \tag{16}$$

Note that $A_j(t, \underline{\omega})$ of definition (8) is not only partitioned into $AD_j^{\langle n\rangle}(t, \underline{\omega})$ and $AL_j^{\langle n\rangle}$ $(t, \underline{\omega})$; $A_j(t, \underline{\omega})$ is also partitioned into the three sets, $D_j^{\langle n\rangle}(t, \underline{\omega})$, $L_j^{\langle n\rangle}(t, \underline{\omega})$, and $Q_j^{\langle n\rangle}(t, \underline{\omega})$. Therefore,

$$m\{A_j(t, \underline{\omega})\} = m\{D_j^{\langle n\rangle}(t, \underline{\omega})\} + m\{L_j^{\langle n\rangle}(t, \underline{\omega})\} + m\{Q_j^{\langle n\rangle}(t, \underline{\omega})\}, \tag{17}$$

(where $m\{Q_j^{\langle n\rangle}(t,\underline{\omega})\}$ is another way of expressing the buffer occupancy, $q_j^{\langle n\rangle}$ $(t,\underline{\omega})$).

Because we are considering work-conserving queueing systems, the amounts of lost class j fluid referred to in Lemma 1 are invariant of which queueing discipline is applied within class j as long as the discipline is work-conserving [14]. We find it convenient to assume that in serving and buffering class j fluid at each time t, system $\langle 2\rangle$ gives strict priority to the fluid in the set, $AD_j^{\langle 1\rangle}(t,\underline{\omega})$, within class j.

To prove (13), we consider the dynamics of class j fluid in $AD_j^{\langle 1\rangle}(t,\underline{\omega})$ that would occur in system $\langle 2\rangle$ if this system had infinite available buffer space for class j fluid. This hypothetical version of system $\langle 2\rangle$, referred to as system $\langle 2'\rangle$, would be identical to system $\langle 2\rangle$ except for having $b_j^{\langle 2'\rangle}(t,\underline{\omega}) = \infty$ for all $t \geq 0$, e.g.,

$$\mu_j^{\langle 2'\rangle}(t,\underline{\omega}) = \mu_j^{\langle 2\rangle}(t,\underline{\omega}), \quad \forall t \geq 0. \tag{18}$$

With no elements of $AD_j^{\langle 1\rangle}(t,\underline{\omega})$ lost in system $\langle 2'\rangle$, $AD_j^{\langle 1\rangle}(t,\underline{\omega})$ is then partitioned into the pair of sets,

$$\tilde{D}_j^{\langle 2'\rangle}(t,\underline{\omega}) := D_j^{\langle 2'\rangle}(t,\underline{\omega}) \cap AD_j^{\langle 1\rangle}(t,\underline{\omega})$$

and

$$\tilde{Q}_j^{\langle 2'\rangle}(t,\underline{\omega}) := Q_j^{\langle 2'\rangle}(t,\underline{\omega}) \cap AD_j^{\langle 1\rangle}(t,\underline{\omega}), \tag{19}$$

so that we have

$$m\{AD_j^{\langle 1\rangle}(t,\underline{\omega})\} = m\{\tilde{D}_j^{\langle 2'\rangle}(t,\underline{\omega})\} + m\{\tilde{Q}_j^{\langle 2'\rangle}(t,\underline{\omega})\}. \tag{20}$$

The main part of our proof of (13) is to show the following proposition.

Proposition 1.

$$m\{D_j^{\langle 1\rangle}(t,\underline{\omega})\} \leq m\{\tilde{D}_j^{\langle 2'\rangle}(t,\underline{\omega})\}, \forall t \geq 0. \tag{21}$$

Proof. For time t such that $m\{\tilde{Q}_j^{\langle 2'\rangle}(t,\underline{\omega})\} = 0$, from (20) and definition (15) we have

$$m\{\tilde{D}_j^{\langle 2'\rangle}(t,\underline{\omega})\} = m\{AD_j^{\langle 1\rangle}(t,\underline{\omega})\} \geq m\{D_j^{\langle 1\rangle}(t,\underline{\omega})\}.$$

For time t such that $m\{\tilde{Q}_j^{\langle 2'\rangle}(t,\underline{\omega})\} > 0$, we define the time,

$$\bar{\sigma} := \sup\{s < t \mid m\{\tilde{Q}_j^{\langle 2'\rangle}(s,\underline{\omega})\} = 0\}, \tag{22}$$

and first show that

$$m\{D_j^{\langle 1\rangle}(\bar{\sigma},\underline{\omega})\} \leq m\{\tilde{D}_j^{\langle 2'\rangle}(\bar{\sigma},\underline{\omega})\}. \tag{23}$$

By definition (22), for any $\varepsilon > 0$, there exists a time,

$$\tau \in [\bar{\sigma} - \varepsilon/\mu, \bar{\sigma}], \tag{24}$$

such that $m\{\tilde{Q}_j^{\langle 2'\rangle}(\tau, \underline{\omega})\} = 0$ (where μ is the finite service rate capacity). By (20) and definition (15),

$$m\big\{\tilde{D}_j^{\langle 2'\rangle}(\tau, \underline{\omega})\big\} = m\big\{AD_j^{\langle 1\rangle}(\tau, \underline{\omega})\big\} \geq m\big\{D_j^{\langle 1\rangle}(\tau, \underline{\omega})\big\}. \tag{25}$$

The amount of class j fluid that is served in system $\langle 1\rangle$ over interval $[\tau, \bar{\sigma}]$ is bounded by μ ($\bar{\sigma}$– τ), i.e.,

$$m\big\{D_j^{\langle 1\rangle}(\bar{\sigma}, \underline{\omega})\big\} - m\big\{D_j^{\langle 1\rangle}(\tau, \underline{\omega})\big\} \leq \mu(\bar{\sigma} - \tau) \leq \varepsilon, \tag{26}$$

where the last inequality is due to (24). Therefore,

$$\begin{aligned} m\big\{\tilde{D}_j^{\langle 2'\rangle}(\bar{\sigma}, \underline{\omega})\big\} &\geq m\big\{\tilde{D}_j^{\langle 2'\rangle}(\tau, \underline{\omega})\big\} && \big(m\big\{\tilde{D}_j^{\langle 2'\rangle}(\cdot, \underline{\omega})\big\}\text{is a nondecreasing function}\big) \\ &\geq m\big\{D_j^{\langle 1\rangle}(\tau, \underline{\omega})\big\} && \text{(by (25))} \\ &\geq m\big\{D_j^{\langle 1\rangle}(\bar{\sigma}, \underline{\omega})\big\} - \varepsilon && \text{(by (26)).} \end{aligned} \tag{27}$$

Then (23) follows from (27) because (27) holds for any positive value, ε. By definition (22), for all $s \in (\bar{\sigma}, t]$, $m\{\tilde{Q}_j^{\langle 2'\rangle}(s, \underline{\omega})\} > 0$, and since this backlogged class j fluid fully utilizes its available service rate,

$$\begin{aligned} m\{\tilde{D}_j^{\langle 2'\rangle}(t, \underline{\omega})\} &= m\big\{\tilde{D}_j^{\langle 2'\rangle}(\bar{\sigma}, \underline{\omega})\big\} + \int_{\bar{\sigma}}^{t} \mu_j^{\langle 2'\rangle}(s, \underline{\omega})\, ds \\ &= m\big\{\tilde{D}_j^{\langle 2'\rangle}(\bar{\sigma}, \underline{\omega})\big\} + \int_{\bar{\sigma}}^{t} \mu_j^{\langle 2\rangle}(s, \underline{\omega})\, ds \quad \text{(by (18))} \\ &\geq m\big\{D_j^{\langle 1\rangle}(\bar{\sigma}, \underline{\omega})\big\} + \int_{\bar{\sigma}}^{t} \mu_j^{\langle 1\rangle}(s, \underline{\omega})\, ds \quad \text{(by (6) and (23))} \\ &\geq m\big\{D_j^{\langle 1\rangle}(t, \underline{\omega})\big\} \text{ (service rate does not exceed available service rate),} \end{aligned}$$

completing the proof of (21). □

Now, for any $t \geq 0$,

$$\begin{aligned} m\big\{\tilde{Q}_j^{\langle 2'\rangle}(t, \underline{\omega})\big\} &= m\big\{AD_j^{\langle 1\rangle}(t, \underline{\omega})\big\} - m\big\{\tilde{D}_j^{\langle 2'\rangle}(t, \underline{\omega})\big\} \quad \text{(by (20))} \\ &\leq m\big\{AD_j^{\langle 1\rangle}(t, \underline{\omega})\big\} - m\big\{D_j^{\langle 1\rangle}(t, \underline{\omega})\big\} \quad \text{(by (21) of Proposition 1)} \\ &= m\{A_j(t, \underline{\omega})\} - m\big\{AL_j^{\langle 1\rangle}(t, \underline{\omega})\big\} - m\big\{D_j^{\langle 1\rangle}(t, \underline{\omega})\big\} \quad \text{(by (11))} \\ &\leq m\{A_j(t, \underline{\omega})\} - m\big\{L_j^{\langle 1\rangle}(t, \underline{\omega})\big\} - m\big\{D_j^{\langle 1\rangle}(t, \underline{\omega})\big\} \quad \text{(by definition (14))} \\ &= m\big\{Q_j^{\langle 1\rangle}(t, \underline{\omega})\big\} \quad \text{(by (17))} \end{aligned}$$

$$
\begin{aligned}
&\le b_j^{\langle 1\rangle}(t,\underline{\omega}) \quad \text{(queue occupancy does not exceed available buffer space)}\\
&\le b_j^{\langle 2\rangle}(t,\underline{\omega}) \quad \text{(by (7))}, \qquad (28)
\end{aligned}
$$

so that the buffer space used by class-j fluid in $AD_j^{\langle 1\rangle}(t,\underline{\omega})$ in system $\langle 2'\rangle$ is never greater than $b_j^{\langle 2\rangle}(t,\underline{\omega})$. In words, the fluid in set $AD_j^{\langle 1\rangle}(t,\underline{\omega})$ is not lost in system $\langle 2\rangle$; to state more rigorously, set

$$\left\{f \in AD_j^{\langle 1\rangle}(t,\underline{\omega}) | f \text{ is lost in system } \langle 2\rangle\right\}$$

has volume 0. Therefore, (13) is proved. Inequality (12) immediately follows from (13) because of (11).

3.4. A corollary to Lemma 1

In this section, we establish the following corollary to Lemma 1.

Corollary 1. For $j \in \{2,3,\dots\}$, and $\underline{\omega} = (\omega_1,\omega_2,\dots,\omega_j)$,

$$\liminf_{T\to\infty} \frac{1}{T} m\left\{L_j^{\langle 2\rangle}(T,\underline{\omega})\right\} \le \liminf_{T\to\infty} \frac{1}{T} m\left\{L_j^{\langle 1\rangle}(T,\underline{\omega})\right\}, \qquad (29)$$

and

$$\limsup_{T\to\infty} \frac{1}{T} m\left\{L_j^{\langle 2\rangle}(T,\underline{\omega})\right\} \le \limsup_{T\to\infty} \frac{1}{T} m\left\{L_j^{\langle 1\rangle}(T,\underline{\omega})\right\}. \qquad (30)$$

Proof. First note that for any class j element, $f \in AL_j^{\langle n\rangle}(t,\underline{\omega})$, either $f \in L_j^{\langle n\rangle}(t,\underline{\omega})$ or

$$f \in QL_j^{\langle n\rangle}(t,\underline{\omega}) := Q_j^{\langle n\rangle}(t,\underline{\omega}) \cap AL_j^{\langle n\rangle}(t,\underline{\omega}), \qquad (31)$$

but not both. Therefore,

$$m\left\{AL_j^{\langle n\rangle}(t,\underline{\omega})\right\} = m\left\{L_j^{\langle n\rangle}(t,\underline{\omega})\right\} + m\left\{QL_j^{\langle n\rangle}(t,\underline{\omega})\right\}, \quad n \in \{1,2\}. \qquad (32)$$

Now, substituting (32) into (12) of Lemma 1 yields

$$
\begin{aligned}
m\left\{L_j^{\langle 2\rangle}(t,\underline{\omega})\right\} &\le m\left\{L_j^{\langle 1\rangle}(t,\underline{\omega})\right\} + m\left\{QL_j^{\langle 1\rangle}(t,\underline{\omega})\right\} - m\left\{QL_j^{\langle 2\rangle}(t,\underline{\omega})\right\}\\
&\le m\left\{L_j^{\langle 1\rangle}(t,\underline{\omega})\right\} + m\left\{QL_j^{\langle 1\rangle}(t,\underline{\omega})\right\}\\
&\le m\left\{L_j^{\langle 1\rangle}(t,\underline{\omega})\right\} + B \qquad (33)
\end{aligned}
$$

(where B is the finite buffer size). Corollary 1 follows immediately from (33). □

3.5. Monotonicity theorem

At this point, we can readily prove the main result of the present paper. Consider a priority-queue that receives fluid of classes $c_1 < c_2 < \cdots$. Then for class c_k, $k \in \{1, 2, \dots\}$, we define

$$L_{c_k}(T; \underline{c}_k) := \{f \mid \text{ element } f \text{ is in class } c_k \text{ and is dropped during } [0, T]\} \qquad (34)$$

where $\underline{c}_k = (c_1, c_2, \dots, c_k)$. (Due to the strict priority, arriving classes $c_{k+1}, c_{k+2}, \dots$ do not affect the set of dropped class c_k fluid elements.) Also, let $L_{c_k}(T, (\omega_{c_1}, \omega_{c_2}, \dots, \omega_{c_k}); \underline{c}_k)$ represent $L_{c_k}(T; \underline{c}_k)$, given $(\omega_{c_1}, \omega_{c_2}, \dots, \omega_{c_k})$ where ω_{c_m} is the arrival process realization of class c_m, $m \in \{1, 2, \dots, k\}$. Now referring again to our previously described system $\langle 1 \rangle$ class j fluid loss set, for any realization, $(\omega_1, \omega_2, \dots, \omega_j)$, system $\langle 1 \rangle$'s cumulative class j loss at time T, $L_j^{\langle 1 \rangle}(T, (\omega_1, \omega_2, \dots, \omega_j))$, can be represented as $L_j(T, (\omega_1, \omega_2, \dots, \omega_j); (1, 2, \dots, j))$; thus, we have the following notational relation between random sets:

$$L_j^{\langle 1 \rangle}(T) = L_j(T; (1, 2, \dots, j)), \qquad (35)$$

where $L_j^{\langle n \rangle}(T)$ is the random set that, upon specification of realization $\underline{\omega}$, yields $L_j^{\langle n \rangle}(T, \underline{\omega})$. Similarly, for system $\langle 2 \rangle$, which receives i $(i < j)$ classes, i.e., classes $1, 2, \dots, i-1, j$, for any realization, $(\omega_1, \omega_2, \dots, \omega_j)$,

$$L_j^{\langle 2 \rangle}(T, (\omega_1, \omega_2, \dots, \omega_j)) = L_j(T, (\omega_1, \omega_2, \dots, \omega_{i-1}, \omega_j); (1, 2, \dots, i-1, j));$$

thus, we have

$$L_j^{\langle 2 \rangle}(T) = L_j(T; (1, 2, \dots, i-1, j)). \qquad (36)$$

We state the monotonicity theorem under the assumption that the fluid loss process in a priority-queue is ergodic in the following sense [3]: there exists the limit,

$$\bar{\ell}_{c_k}(\underline{c}_k) := \lim_{T \to \infty} E\left[\frac{1}{T} m\{L_{c_k}(T; \underline{c}_k)\}\right], \text{ and} \qquad (37)$$

$$\frac{1}{T} m\{L_{c_k}(T; \underline{c}_k)\} \to \bar{\ell}_{c_k}(\underline{c}_k) \text{ as } T \to \infty, \text{ with probability one;} \qquad (38)$$

i.e., for almost every [21] $(\omega_{c_1}, \omega_{c_2}, \dots, \omega_{c_k})$, $\frac{1}{T} m\{L_{c_k}(T, (\omega_{c_1}, \omega_{c_2}, \cdots, \omega_{c_k}); \underline{c}_k)\} \to \bar{\ell}_{c_k}(\underline{c}_k)$ as $T \to \infty$.

For any $k \in \{1, 2, \dots\}$, and considering the class vector, $\underline{c}_k = (1, 2, \dots, k)$, we define

$$\bar{\ell}_k := \bar{\ell}_k((1, 2, \dots, k)). \qquad (39)$$

Then, because all classes of arrival processes are statistically independent, identically distributed random processes,

$$\bar{\ell}_k = \bar{\ell}_{c_k}(\underline{c}_k) \qquad (40)$$

for any k-vector, $\underline{c}_k = (c_1, c_2, \dots, c_k)$, such that $c_1 < c_2 < \cdots < c_k$.

Theorem 1. For the queueing system model of Section 2 (fluid queue with strict priority) receiving fluid of classes 1, 2, ..., and positive integers, $i < j$, we have $\bar{\ell}_i \leq \bar{\ell}_j$, assuming ergodicity as in (38).

Proof. Let $i, j \in \{1, 2, \ldots\}$ such that $i < j$. Then

$$\begin{aligned}
\bar{\ell}_i &= \bar{\ell}_j((1, 2, \ldots, i-1, j)) \quad \text{(by (40) with } \underline{c}_i := (1, 2, \ldots, i-1, j))\\
&= \lim_{T\to\infty} \frac{1}{T} m\{L_j(T; (1, 2, \ldots, i-1, j))\} \quad \text{(with probability one by (38))}\\
&= \lim_{T\to\infty} \frac{1}{T} m\{L_j^{\langle 2\rangle}(T)\} \quad \text{(by (36))}\\
&\leq \lim_{T\to\infty} \frac{1}{T} m\{L_j^{\langle 1\rangle}(T)\} \quad \text{(by Corollary 1)}\\
&= \lim_{T\to\infty} \frac{1}{T} m\{L_j(T; (1, 2, \ldots, j))\} \quad \text{(by (35))}\\
&= \bar{\ell}_j((1, 2, \ldots, j)) \quad \text{(with probability one by (38))}\\
&= \bar{\ell}_j \quad \text{(by (40) with } \underline{c}_j := (1, 2, \ldots, j)).
\end{aligned}$$

□

4. Discussions

The fluid queueing model facilitates the mathematical analysis while providing good physical intuition about many systems; e.g., [1,2,4–13,15–20,22–26]. In this paper, we proved a rather intuitive proposition that for multiple streams with statistically independent, identically distributed random arrival processes, a stream with a higher priority has a lower average loss rate than a stream with a lower priority. We also observe that analysis of a fluid queue at times requires mathematical artifices that are not seen in the analyses of discrete-event queues.

Acknowledgments

The authors would like to thank an Associate Editor and an anonymous reviewer for their helpful comments.

References

[1] E. Altman, T. Jimenez and G. Koole, On optimal call admission control in resource-sharing system, IEEE Trans. Commun. 49(9) (2001) 1659–1668.

[2] D. Bertsimas, D. Gamarnik and J. Tsitsiklis, Stability conditions for multiclass fluid queueing networks, IEEE Trans. Automat. Control 41(11) (1996) 1618–1631.

[3] D. Bertsekas and R. Gallager, *Data Networks*, 2nd Ed. (Prentice Hall, 1992).

[4] N. Christin, J. Liebeherr and T.F. Abdelzaher, A quantitative assured forwarding service, in: *Proc. IEEE INFOCOM 2002*, Vol. 2 (2002) pp. 864–873.

[5] B.D. Choi, B.C. Shin, K.B. Choi, D.H. Han and J.S. Jang, Priority queue with two state Markov modulated arrivals, in: *Proc. IEEE ICC '96*, Vol. 2 (1996) pp. 1055–1059.

[6] J. Choe and N.B. Shroff, A central-limit-theorem-based approach for analyzing queue behavior in high-speed networks, IEEE/ACM Trans. Networking 6(5) (1998) 659–671.

[7] J. Choe and N.B. Shroff, On the supremum distribution of integrated stationary Gaussian processes with negative linear drift, Adv. in Appl. Probab. 31(1) (1999) 135–157.

[8] A. Elwalid and D. Mitra, Analysis, approximations and admission control of a multi-service multiplexing system with priorities, in: *Proc. IEEE INFOCOM '95*, Vol. 2 (1995) pp. 463–472.

[9] A.I. Elwalid and D. Mitra, Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic, in: *Proc. IEEE INFOCOM '92*, Vol. 1 (1992) pp. 415–425.

[10] P. Fonseca, J.M. Pitts and L.G. Cuthbert, Exact fluid-flow analysis of single ON/OFF source feeding an ATM buffer with space priority, Electronics Letters 31(13) (1995) 1028–1029.

[11] B. Hajek, Large bursts do not cause instability, IEEE Trans. Automat. Control 45(1) (2000) 116–118.

[12] F.P. Kelly, Models for a self-managed Internet, Philosoph. Trans. Royal Soc. London A358 (2000) 2335–2348.

[13] C. Knessl and C. Tier, A simple fluid model for servicing priority traffic, IEEE Trans. Automat. Control 45(6) (2001) 909–914.

[14] L. Kleinrock, *Queueing Systems Volume II: Computer Applications* (Wiley Interscience, 1975).

[15] J. Liebeherr, S.D. Patek and A. Burchard, Statistical per-flow service bounds in a network with aggregate provisioning, in: *Proc. IEEE INFOCOM 2003*, Vol. 3 (2003) pp. 1680–1690.

[16] S.H. Low and P.P. Varaiya, Burst reducing servers in ATM networks, Queueing Systems 20 (1995) 61–84.

[17] S.H. Low and P.P. Varaiya, A new approach to service provisioning in ATM networks, IEEE/ACM Trans. Networking 1(5) (1993) 547–553.

[18] Y. Liu and W. Gong, On fluid queueing system with strict priority, in: *Proc. IEEE Conf. on Decision and Control*, Vol. 2 (2001) pp. 1923–1928.

[19] D.C. Lee, Worst-case fraction of CBR teletraffic unpunctual due to statistical multiplexing, IEEE/ACM Trans. Networking 4(1) (1996) 98–105.

[20] A. Narayanan and V.G. Kulkarni, First passage times in fluid models with an application to two priority fluid systems, in: *Proc. IEEE IPDS '96* (1996) pp. 166–175.

[21] H.L. Royden, *Real Analysis*, 3rd Ed. (Prentice-Hall, 1988).

[22] S. Shakkottai and R. Srikant, Deterministic fluid models of congestion control in high-speed networks, in: *Proc. Winter Simulations Conf.*, Arlington, VA (December, 2001).

[23] B. Sericola and B. Tuffin, A fluid queue driven by a Markovian queue, Queueing Systems 31(3/4) (1999) 253–264.

[24] L. Tassiulas and A. Ephremides, Jointly optimal routing and scheduling in packet radio networks, IEEE Trans. Inform. Theory 38(1) (1992) 165–168.

[25] D.N.C. Tse, R.G. Gallager and J.N. Tsitsiklis, Statistical multiplexing of multiple time-scale Markov streams, IEEE J. Selected Areas in Commun. 13(6) (1995) 1028–1038.

[26] J. Zhang, Performance study of Markov modulated fluid flow models with priority traffic, in: *Proc. IEEE INFOCOM '93*, Vol. 1 (1993) pp. 10–17.