Push Forward Link-Level Scheduling for Network-Wide Performance

Leandros Tassiulas, *Member, IEEE*

***Abstract*— A virtual circuit network with arbitrary topology is considered. The traffic streams follow prespecified routes, different in general for each stream, to reach their destination. A fluid traffic model is adopted and a processor sharing service discipline is considered. A policy is proposed for setting adaptively the fractions of the transmission capacity, which is allocated to the different traffic streams in the processor sharing discipline at each link. The amount of traffic arrived at the originating node of each link is measured for each stream. The fraction of the link capacity allocated to each stream is set to be proportional to the measured traffic. The traffic is measured continuously and the fractions are updated regularly based on the most recent traffic measurements. It is shown that eventually, the transmission capacity allocated to each stream converges to a quantity proportional to the average rate of the stream. Hence, if the capacity condition is satisfied, sufficient fractions of the capacity are allocated at each link for each stream. End-to-end performance guarantees are provided, if the traffic is regulated. The policy is distributed since each link adjusts the service fractions based on observations of the traffic arriving at its originating node only. Furthermore, it is adaptive since no information on the traffic characteristics is needed for the application of the policy.**

I. Introduction

IN THIS PAPER, we focus on the congestion control functions of a packet switched network. The model of a virtual circuit (VC) network is considered. It consists of a set of switching nodes, a set of links connecting certain pairs of the switching nodes and a set of information streams (Fig. 1). An information stream enters the network at its originating node and after crossing a sequence of links (its path), exits the network at its destination node. Nonblocking switching is assumed at each network node. When a packet enters a node, either from outside or arriving from another network node, it is queued in front of the output link that it is going to cross next. If the node is the eventual destination of the packet then the latter departs from the network. In the originating node of each link there are packets from the different information streams that go through the link waiting to be served. The sharing of the link service capacity among the different traffic streams and the associated congestion control problem is the subject of this paper.

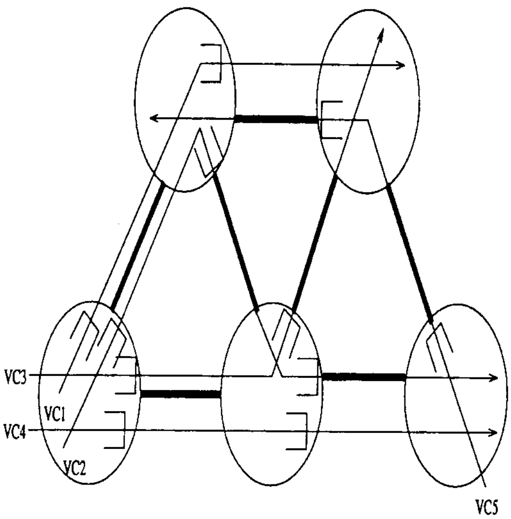

Fig. 1. The queueing model of a virtual circuit network.

A fluid traffic model is considered. The traffic is infinitely divisible and the server may switch from stream to stream in infinitesimal time. The link capacity is allocated in a processor sharing fashion. The congestion control scheme is specified by the fractions of the capacity allocated to the different traffic streams. These may vary with time. A policy is proposed, which adaptively adjusts the capacity fractions, such that eventually adequate capacity is allocated to each traffic stream at each link. It is called the push forward (PF) policy and acts as follows. At the originating node of each link l, the arriving traffic of each stream j that crosses l is monitored and a measurement of the total amount of traffic $Q_j^l(t)$ that has arrived until time t is kept. At certain time instants, the update times, the fractions of the service capacity that correspond to each stream j through link l are updated to be proportional to the most recent measurements of the arriving traffic. The sequence of update times can be fairly arbitrary. It is shown that the capacity fractions allocated to the different streams that cross l converge to values proportional to the rates of the streams. This holds under the condition that the average rates exist and the capacity condition is satisfied at each link. Hence, the existence of the long-run average arrival rate of each stream guarantees that the output rate will be equal to the input rate. If the arrivals are regulated and the stronger condition that the arrival stream j has bounded burstiness holds, then it is guaranteed that the backlog of the stream j at every link that the stream goes through, is bounded as well. These results hold for every arrival sample path that satisfies certain deterministic conditions.

Manuscript received February 16, 1994; revised December 21, 1995; approved by IEEE/ACM Transactions on Networking Editor M. Sidi. This work was supported in part by the National Science Foundation under Grants NCR-9211417, NCR-9406415, the AFOSR under Grant 95-1-0061, and the New York State Center for Advanced Technology in Telecommunications, Polytechnic University.

The author is with the Department of Electrical Engineering, University of Maryland, College Park, MD 20742 USA.

Publisher Item Identifier S 1063-6692(96)04490-1.

The problem of congestion control for providing performance guarantees in communication networks has been studied extensively lately and several new methods of traffic modeling and analysis have been proposed and investigated by several authors [1], [3], [6]–[8], [10], [12]. The generalized processor sharing (GPS) scheme, proposed for congestion control by Parekh and Gallager [8], guarantees a certain amount of bandwidth to each traffic stream. The PF policy provides a method for adjusting the fractions of allocated bandwidth in a GPS scheme adaptively, without knowledge of the traffic characteristics. Only the traffic arriving at the originating node of a link needs to be observed for the adjustment of the bandwidth allocations of the link. Hence, the policy can be implemented in a distributed fashion.

The paper is organized as follows. In Section II, the network model is defined, the traffic assumptions are given, and the scheduling discipline is discussed. In Section III, the congestion control policy is specified. In Section IV, the network is analyzed under the assumption that the long-run average rates exist. In Section V, the network is analyzed under the additional assumption that the arrival streams have bounded burstiness. Section VI contains some concluding remarks.

II. The Network Model and the Traffic Assumptions

The network has L links and a total number of J traffic streams. The streams may be routed arbitrarily through the network. The set of links that a particular traffic stream j goes through is denoted by $L(j)$ and the kth link, in the order that they are crossed by stream j is denoted by $l_j(k)$. The number of links in $L(j)$ is denoted by K_j. The set of all traffic streams that gó through link l is denoted by $V(l)$. Denote by $p_j(l)$ the link through which stream j arrives at the originating node of link l. By convention, if stream j enters the network at the originating node of link l then $p_j(l) = 0$. Fig. 2 clarifies the notation defined above.

The traffic of each session j arrives at its entry node of the network that need not be the same for all sessions. The arrivals of stream j are specified by the sequence $\{(\tau_n^j, \sigma_n^j)\}_{n=1}^{\infty}$ where τ_n^j, σ_n^j are the arrival time and the information length of the nth packet of the stream, respectively. An alternative representation of the arrival process is given by the instantaneous rate $a_j(t)$ with which information of stream j is entering the network at time t. We assume $a_j(t) = 0$ for $t < 0$. The amount of information arrived in the network within the time interval $[t_1, t_2)$ is equal to $\int_{t_1}^{t_2} a_j(\tau)\, d\tau$. Clearly, we have

$$\int_0^{\tau_n^j} a_j(\tau)\, d\tau = \sum_{l=1}^{n-1} \sigma_l^j, \quad \forall n > 0.$$

In this paper, we adopt the representation of the traffic by the instantaneous arrival rate. The following conditions about the arrival streams are assumed to hold throughout the paper.

C1: The packet length is bounded

$$\sigma_n^j \leq c, j = 1, \cdots, J, \qquad n = 1, 2, \cdots$$

C2: The long-run average arrival rates exist for all streams

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t a_j(\tau)\, d\tau = a_j, \qquad j = 1, \cdots, J.$$

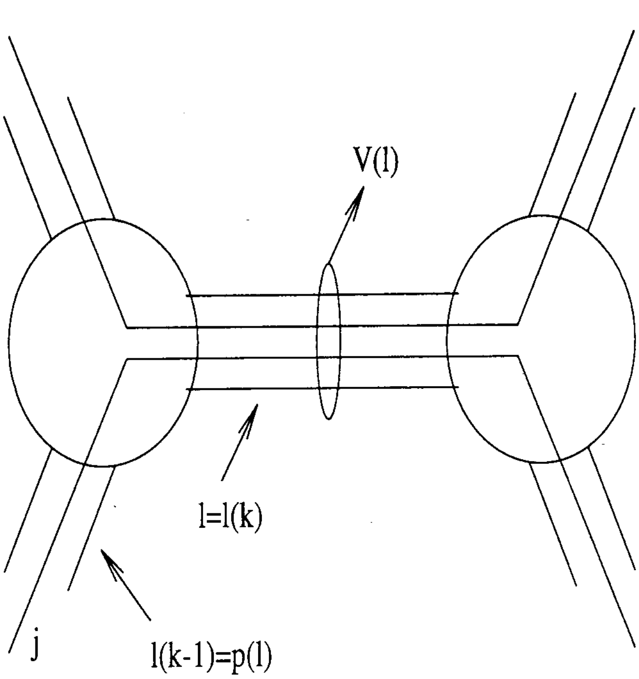

Fig. 2. The notation is illustrated in this figure. The kth link crossed by stream j is denoted by $l_j(k)$. The set of streams crossing link l is denoted by $V(l)$. The link crossed by stream j before this goes through l is denoted by $p_j(l)$.

C3: For all links we have

$$\sum_{j \in V(l)} a_j \leq C_l, \qquad l = 1, \cdots. L$$

where C_l is the transmission capacity of link l.

C4: There are no instantaneous packet arrivals, that is

$$a_j(t) < \infty, \qquad t \geq 0, \quad j = 1, \cdots, J.$$

Note that no specific statistical model is considered for the traffic. There is just a set of conditions that should hold for every sample path of the arrival stream. The results obtained in the rest of this paper hold for every arrival sample path that satisfies the imposed conditions and not just for statistical averages. Conditions C1–C4 imply some weak stability properties of the network under the PF policy. Stronger stability properties and bounds on the backlog are obtained in Section V assuming additional constraints on the burstiness of the arrival streams.

III. Congestion Control and the Push-Forward Policy

The link transmission capacity may be allocated to more than one stream simultaneously, in a processor sharing fashion. The transmission of a packet may be initiated only after it has completely arrived at the originating node of the link. The traffic of each stream may be transmitted with any rate as long as the total transmission rate of all the streams through the link does not exceed the total link transmission capacity at any time. The instantaneous rate with which information of stream j is flowing at time t through link l is denoted by $\mu_j^l(t)$. By convention, we have $\mu_j^0(t) = a_j(t)$. If the transmission of the nth packet of stream j through link l ends at time t and of the packet $n+1$ at t', then

$$\int_t^{t'} \mu_j^l(\tau)\, d\tau = \sigma_{n+1}^j.$$

Clearly, the aggregate transmission rate through link l should satisfy the capacity constraint

$$\sum_{j\in V(l)} \mu_j^l(t) \le C_l, t \ge 0. \tag{1}$$

Let $X_j^l(t)$ be the amount of information of stream j in the originating node of link l at time t. It includes the amount of information in stream j packets waiting for transmission, the portion of the stream j packet under transmission through l that has not left the node yet, and the portion of stream j packet under transmission through link $p_j(l)$ that has already arrived in the node. Clearly, we have

$$X_j^l(t_2) - X_j^l(t_1) = \int_{t_1}^{t_2} (\mu_j^{p_j(l)}(\tau) - \mu_j^l(\tau))\, d\tau, \quad t_2 > t_1 \ge 0. \tag{2}$$

Condition 1 implies that the streams from one node to the other contain no impulses and together with the condition C4 implies that the backlog of any traffic stream at any node cannot have jumps as a function of time. The rates $\mu_j^l(t)$ are adjusted by the congestion control scheme. *Congestion control* policy is any rule for determining a set of transmission rates $\{\{\mu_j^l(t), t \ge 0\}, j \in V(l), l = 1, \cdots, L\}$.

One congestion control policy that provides guaranteed bandwidth to the individual traffic streams is GPS. In that scheme, "weights" f_j^l are specified for each stream $j \in V(l)$ for every link l. They represent the capacity fraction of link l allocated to stream j. Then, at every time instant t, the rates $\mu_j^l(t)$, $\mu_k^l(t)$ of any two sessions j, k with nonzero backlogs, are such that $\mu_j^l(t)/f_j^l = \mu_k^l(t)/f_k^l$. The scheme has been studied extensively in both the single node [8] and the network case [9] and it was shown that it can provide performance guarantees when the weights are selected appropriately. In GPS, the weights are preallocated based on the traffic characteristics of the streams, and remain fixed thereafter. The preallocation should be done such that fair portions of the capacity are allocated to the individual streams.

A. The Push-Forward Policy

In this paper, it is assumed that the capacity fractions allocated to the different streams may change with time. The problem of adjusting them adaptively based on the observed traffic characteristics is considered. The PF policy makes the adaptive allocation such that, eventually, the capacity fractions converge to values larger than the traffic loads. The service fractions of each stream for link l may change only at the *update time instants*, $\{t^l(n)\}_{n=1}^{\infty}$ of link l. That is, the capacity allocation scheme is piecewise constant. It is denoted as

$$\{\{(t^l(n), f_n^{lj})\}_{n=0}^{\infty}, j \in V(l)\}$$

with the interpretation that

$$f_j^l(t) = f_n^{lj}, t^l(n+1) > t \ge t^l(n).$$

The update times of each link can be any sequence such that $\lim_{n\to\infty} t^l(n) = \infty$. This assumption implies that the capacity of a link cannot be reallocated an infinite number of times during a finite time interval and it is introduced for technical reasons only. For the results obtained in this paper, the update sequences for different links do not need to satisfy jointly any constraint, and can be completely asynchronous.

The fraction f_n^{lj} of the link l capacity, allocated to stream $j \in V(l)$ during the period $[t^l(n), t^l(n+1))$ is

$$f_n^{lj} = \frac{Q_j^l(t^l(n))}{\sum_{m\in V(l)} Q_m^l(t^l(n))} \tag{3}$$

where $Q_j^l(t)$ is the total amount of stream j traffic that arrived at the originating node of link l until time t. That is

$$Q_j^l(t) = \int_0^t \mu_j^{p_j(l)}(\tau)\, d\tau + X_j^l(0)$$

where $X_j^l(0)$ is the initial backlog of stream j at the originating node of link l. The transmission rates $\mu_j^l(t)$ should be such that

$$\mu_j^l(t) \ge f_j^l(t) C_l, \qquad j \in V(l), l = 1, \cdots, L \tag{4}$$

if at time t there is a stream j packet in the originating node of link l. Of course (1) should be satisfied. The capacity fractions f_n^{lj} under the PF policy, converge in some sense to those of the rate proportional processor sharing policy, considered by Parekh and Gallager in [9].

Note that, strictly speaking, the PF policy as defined above is a class of policies and not a single policy, since there is more than one way to select $\mu_j^l(t)$'s, such that (4) is satisfied. The results obtained later hold for any policy of this class. So they will not be distinguished in the following and they will be referred to collectively as the PF policy.

IV. Convergence of the Bandwidth Allocation and Existence of the Output Rates

In this section, it is shown that under the PF policy, the average output rate from any link l of each stream $j \in V(l)$ is equal to the average rate a_j of the stream. Furthermore, the service fraction f_n^{lj} converges to the relative load of j, among the streams in $V(l)$. These results are stated in Theorems 1 and 2, respectively, and they follow from the conditions C1–4 on the arrival streams.

Theorem 1: Under the PF policy, if conditions C1–4 hold, then the average transmission rate of any stream j through any link $l \in L(j)$ exists and is equal to the arrival rate of the stream

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \mu_j^l(\tau)\, d\tau = a_j. \tag{5}$$

The proof of the theorem will follow after two lemmas. The following lemma shows that if the average rate with which stream j arrives at the originating node of link l exists and is equal to the arrival rate of the stream a_j, then the service fraction allocated to stream j at link l under the PF policy will eventually be larger than or equal to the average rate a_j.

Lemma 1: If condition C2 holds and at the originating node of link l, the traffic of stream j satisfies the condition

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \mu_j^{p_j(l)}(\tau)\, d\tau = a_j \tag{6}$$

and also

$$\sum_{j\in V(l)} a_j < C_l \tag{7}$$

then

$$\lim \inf_{n\to\infty} f_n^{l,j} > \frac{a_j}{C_l}. \tag{8}$$

While, if

$$\sum_{j\in V(l)} a_j = C_l \tag{9}$$

then

$$\lim \inf_{n\to\infty} f_n^{l,j} \geq \frac{a_j}{C_l}. \tag{10}$$

Proof: Notice that for the stream j in link l, we have

$$Q_j^l(t^l(n)) = \int_0^{t^l(n)} \mu_j^{p_j(l)}(\tau)\, d\tau + X_j^l(0) \tag{11}$$

and for any stream $k \in V(l)$, we have

$$Q_k^l(t^l(n)) \leq \int_0^{t^l(n)} a_k(\tau)\, d\tau + X_k^l(0). \tag{12}$$

From (11) and (12) we get

$$\begin{aligned} f_n^{l,j} &= \frac{Q_j^l(t^l(n))}{\sum\limits_{k\in V(l)} Q_k^l(t^l(n))} \\ &\geq \frac{\displaystyle\int_0^{t^l(n)} \mu_j^{p_j(l)}(\tau)\, d\tau + X_j^l(0)}{\sum\limits_{k\in V(l)} \left(\int_0^{t^l(n)} a_k(\tau)\, d\tau + X_k^l(0)\right)}. \end{aligned} \tag{13}$$

By taking the limits in (13) and from condition C2, we get

$$\begin{aligned} &\lim \inf_{n\to\infty} f_n^{l,j} \\ &\geq \lim \inf_{n\to\infty} \frac{\displaystyle\int_0^{t^l(n)} \mu_j^{p_j(l)}(\tau)\, d\tau + X_j^l(0)}{\sum\limits_{k\in V(l)} \left(\displaystyle\int_0^{t^l(n)} a_k(\tau)\, d\tau + X_k^l(0)\right)} \\ &\geq \frac{\lim \inf\limits_{n\to\infty} \dfrac{1}{t^l(n)} \left(\displaystyle\int_0^{t^l(n)} \mu_j^{p_j(l)}(\tau)\, d\tau + X_j^l(0)\right)}{\lim\limits_{n\to\infty} \dfrac{1}{t^l(n)} \sum\limits_{k\in V(l)} \left(\displaystyle\int_0^{t^l(n)} a_k(\tau)\, d\tau + X_k^l(0)\right)} \\ &= \frac{a_j}{\sum\limits_{k\in V(l)} a_k}. \end{aligned} \tag{14}$$

From (7) and (14), (8) follows while (10) follows from (9) and (14). ◇

The following lemma shows that if the long-run average rate of stream j at the input of link l exists, then the long-run average rate of the stream j at the output of link l exists as well and is equal to that of the input.

Lemma 2: If conditions C1–4 hold and at the originating node of link l the traffic of stream $j \in V(l)$ satisfies the condition

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \mu_j^{p_j(l)}(\tau)\, d\tau = a_j \tag{15}$$

then the output rate of stream j from link l exists and is

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \mu_j^l(\tau)\, d\tau = a_j. \tag{16}$$

Proof: For every stream j in every link $l \in L(j)$, we have

$$\frac{1}{t} \int_0^t \mu_j^l(\tau)\, d\tau \leq \frac{1}{t} \int_0^t \mu_j^{p_j(l)}(\tau)\, d\tau + \frac{1}{t} X_j^l(0) \tag{17}$$

from which we get using (15)

$$\begin{aligned} \lim \sup_{t\to\infty} \frac{1}{t} \int_0^t \mu_j^l(\tau)\, d\tau &\leq \lim \sup_{t\to\infty} \left\{ \frac{1}{t} \int_0^t \mu_j^{p_j(l)}(\tau)\, d\tau \right. \\ &\qquad \left. + \frac{1}{t} X_j^l(0) \right\} = a_j. \end{aligned} \tag{18}$$

Given (18), in order to prove the lemma, it is enough to show that

$$\lim \inf_{t\to\infty} \frac{1}{t} \int_0^t \mu_j^l(\tau)\, d\tau \geq a_j$$

or, equivalently, that

$$\lim \inf_{t\to\infty} \frac{1}{t} \int_0^t \mu_j^l(\tau)\, d\tau \geq a_j - \epsilon, \quad \forall \epsilon > 0. \tag{19}$$

We will show (19) by contradiction. Assume that for some $\epsilon_0 > 0$, we have

$$\lim \inf_{t\to\infty} \frac{1}{t} \int_0^t \mu_j^l(\tau)\, d\tau < a_j - \epsilon_0 \tag{20}$$

which implies that there exists a sequence $t_n, n = 1, 2, \cdots$ such that

$$\lim_{n\to\infty} t_n = \infty \tag{21}$$

and

$$\frac{1}{t_n} \int_0^{t_n} \mu_j^l(\tau)\, d\tau < a_j - \epsilon_0. \tag{22}$$

From Lemma 1, we have

$$\lim \inf_{n\to\infty} f_n^{l,j} \geq \frac{a_j}{C_l} \tag{23}$$

which together with (15) implies that there exists t^* such that

$$f_j^l(t) \geq \frac{(a_j - \epsilon_0/2)}{C_l}, \qquad t \geq t^* \tag{24}$$

and

$$\frac{1}{t} \int_0^t \mu_j^{p_j(l)}(\tau)\, d\tau \geq a_j - \frac{\epsilon_0}{2} + \frac{c}{t}, \qquad t \geq t^*. \tag{25}$$

It is claimed that conditions (20) and (24) imply that the backlog $X_j^l(t)$ becomes less than the maximum packet length c infinitely often or, in other words, that there exists a sequence $t'_n, n = 1, 2, \cdots$ such that

$$\lim_{n\to\infty} t'_n = \infty \tag{26}$$

and

$$X_j^l(t'_n) < c. \tag{27}$$

This is shown in the following by contradiction. If that was not true, then there should have been a time $\tilde{t}$ such that

$$X_j^l(t) \ge c, \qquad t \ge \tilde{t}. \tag{28}$$

Clearly, we would have had

$$\mu_j^l(t) \ge f_j^l(t) C_l \ge a_j - \epsilon_0/2, \qquad t \ge \max\{t^*, \tilde{t}\} \tag{29}$$

therefore

$$\frac{1}{t}\int_0^t \mu_j^l(\tau)\, d\tau \ge \frac{1}{t}\int_{\tilde{t}}^t (a_j - \epsilon_0/2)\, d\tau, \qquad t > \tilde{t} \tag{30}$$

from which, we get

$$\lim_{t\to\infty}\inf \frac{1}{t}\int_0^t \mu_j^l(\tau)\, d\tau \ge a_j - \epsilon_0/2 \tag{31}$$

which contradicts (20).

After showing the existence of the sequences t_n and t'_n we continue to show that (20) leads to contradiction. Select a t'_n larger than t^* and a t_k larger than t'_n. Let $\hat{t}$ be defined as

$$\hat{t} = \sup\{t\colon t \le t_k, X_j^l(t) < c\}.$$

Notice that since $t^* < t'_n < t_k$, the time $\hat{t}$ is well defined and $t'_n \le \hat{t} \le t_k$. We can write

$$\frac{1}{t_k}\int_0^{t_k} \mu_j^l(\tau)\, d\tau = \frac{1}{t_k}\int_0^{\hat{t}} \mu_j^l(\tau)\, d\tau + \frac{1}{t_k}\int_{\hat{t}}^{t_k} \mu_j^l(\tau)\, d\tau. \tag{32}$$

By the definition of $\hat{t}$ we have that $X_j^l(t) \ge c$ for t in the interval $(\hat{t}, t_k)$, therefore, stream j is transmitted with rate $\mu_j^l(t) \ge a_j - \epsilon_0/2$ because of (24). For the second term of the sum in the right-hand side of (32), and since $\hat{t} > t^*$, we have

$$\frac{1}{t_k}\int_{\hat{t}}^{t_k} \mu_j^l(\tau)\, d\tau \ge \frac{t_k - \hat{t}}{t_k}(a_j - \epsilon_0/2). \tag{33}$$

For the first term of the sum in the right-hand side of (32), we have

$$\int_0^{\hat{t}} \mu_j^l(\tau)\, d\tau + c = \int_0^{\hat{t}} \mu_j^{p_j(l)}(\tau)\, d\tau + X_j^l(0)$$

which implies

$$\frac{1}{\hat{t}}\int_0^{\hat{t}} \mu_j^l(\tau)\, d\tau \ge \frac{1}{\hat{t}}\int_0^{\hat{t}} \mu_j^{p_j(l)}(\tau)\, d\tau - \frac{c}{\hat{t}}. \tag{34}$$

From the definition of $\hat{t}$ and (25), (34) gives

$$\frac{1}{\hat{t}}\int_0^{\hat{t}} \mu_j^l(\tau)\, d\tau \ge a_j - \epsilon_0/2 \Rightarrow$$
$$\frac{1}{t_k}\int_0^{\hat{t}} \mu_j^l(\tau)\, d\tau \ge \frac{\hat{t}}{t_k}(a_j - \epsilon_0/2). \tag{35}$$

In view of (33) and (35), (32) gives

$$\frac{1}{t_k}\int_0^{t_k} \mu_j^l(\tau)\, d\tau \ge a_j - \epsilon_0/2 \tag{36}$$

which contradicts (22). ◇

We proceed now to the proof of the Theorem 1.

Proof of Theorem 1: For each stream j, we show by induction that (5) is satisfied for all links $l \in L(j)$.

For the first link $l_j(1)$, that stream j traverses when it enters the network, we have $\mu_j^{p_j(l)}(t) = a_j(t)$, therefore, (15) holds from condition C2 and Lemma 2 implies readily (5).

Assume that (5) holds for $l = l_j(k)$. Notice that relation (5) for $l = l_j(k)$ is identical to (15) for $l = l_j(k+1)$. Therefore, Lemma 2 implies (5) for $l = l_j(k+1)$ as well. The proof is concluded by induction. ◇

The following theorem strengthens the convergence results (8) and (10) showing that the service fractions of each stream converge under the PF policy. Notice that the results (8) and (10) on the convergence of the bandwidth fractions helped us to show the convergence of the average rates for the streams at every network stage, that is, Theorem 1. The latter is used to show the convergence of the service fractions.

Theorem 2: Under the PF policy, if conditions C1–4 hold, then the service fraction allocated to every information stream at every link j *converges to*

$$\lim_{n\to\infty} f_n^{lj} = \frac{a_j}{\sum\limits_{k\in V(l)} a_k}. \tag{37}$$

Proof: Notice that

$$Q_j^l(t^l(n)) = X_j^l(0) + \int_0^{t^l(n)} \mu_j^{p_j(l)}(\tau)\, d\tau, \qquad j \in V(l). \tag{38}$$

The theorem follows readily from (38), the definition of f_n^{lj} and Theorem 1. ◇

The fact that the output rate of a stream is equal to the input rate is a rather weak stability condition since it does not exclude the possibility of unbounded backlogs inside the network. Under the conditions C1–4 though, no stronger stability condition holds. Even in the special case of a single queue, the condition that the arrival rate is smaller than the service rate does not guarantee bounded queue lengths without additional conditions on the arrival streams. In the next section, we study the backlogs in the network nodes under constraints on the burstiness of the arrival streams.

V. Burstiness Control

When the arrival streams have bounded burstiness, then it is expected stronger stability properties are satisfied by the network than the mere existence of the output rates. If the rate proportional processor sharing is employed and the burstiness is bounded, then it is shown by Parekh and Gallager [9] that the backlogs in the network nodes are bounded. The rate proportional processor sharing is, in some sense, the limiting policy of the adaptive PF policy proposed here, so it is expected that after an initial transient period during which the backlogs may fluctuate unpredictably, they will eventually settle down, below certain bounds independent of the initial conditions. This is shown in the rest of this section.

The burstiness b_j of a traffic stream $a_j(t)$ with long-run average a_j was defined by Cruz in [2] as

$$b_j = \sup_{t_2 \geq t_1 \geq 0} \left\{ \int_{t_1}^{t_2} a_j(\tau)\, d\tau - (t_2 - t_1) a_j \right\}. \tag{39}$$

A traffic stream at the output of a traffic regulator (leaky bucket) and before it enters the network has finite burstiness of size determined by the parameters of the regulator. In the analysis of the backlogs during the operation of the congestion control policy, there is a need to characterize the burstiness of the traffic streams within the network when the capacity fractions will converge and after the initial transient period. To facilitate this task, the notion of *eventual burstiness* is defined in the following. The arrival stream j with average rate a_j has eventual burstiness $\hat{b}_j$ if

$$\lim_{t \to \infty} \sup_{t_2 \geq t_1 \geq t} \left\{ \int_{t_1}^{t_2} a_j(\tau)\, d\tau - a_j(t_2 - t_1) \right\} = \hat{b}_j. \tag{40}$$

Roughly speaking, eventual burstiness is the maximum burstiness of a traffic stream after a sufficiently large transient period.

The following theorem characterizes the eventual burstiness of a traffic stream after it crosses each link as it goes through the network. The maximum backlog in each node follows easily from that characterization.

Theorem 3: If conditions C1–2 hold, the arrival stream j has eventual burstiness $\hat{b}_j$ and for all links $l \in L(j)$ it holds

$$\sum_{j \in V(l)} a_j < C_l \tag{41}$$

then the stream $\{\mu_j^{l_j(k)}(t), t \geq 0\}$ has eventual burstiness less than or equal to $\hat{b}_j + kc$ for $k = 1, \cdots, K_j$.

The proof of Theorem 3 will follow after the following lemma which characterizes the eventual burstiness of the output of a stream as it goes through a link, given a constraint on the eventual burstiness of the input stream.

Lemma 3: If the traffic stream $\{\mu_j^{p_j(l)}(t), t \geq 0\}$ has eventual burstiness less than or equal to $\hat{b}_j$, the conditions C1–2 hold and

$$\sum_{j \in V(l)} a_j < C_l \tag{42}$$

then the stream $\{\mu_j^l(t), t \geq 0\}$ has eventual burstiness less than or equal to $\hat{b}_j + c$.

Proof: Under the conditions of Lemma 3, we have from Theorem 2 that

$$\lim_{n \to \infty} f_n^{lj} = \frac{a_j}{\sum_{k \in V(l)} a_k}. \tag{43}$$

From (42) and (43), we understand that there exist $a^* > a_j$ and t^* such that

$$f_j^l(t) C_l \geq a^*, \qquad t \geq t^*. \tag{44}$$

At all times $t \geq t^*$, the following holds. If there is not a complete packet at the originating node of link l, then the rate $\mu_j^l(t)$ is equal to zero, otherwise it is greater than or equal to a^*.

We argue that there is a time $t^{**} \geq t^*$ such that $X_j^l(t^{**}) < c$. If that was not true and $X_j^l(t) \geq c$ for all $t \geq t^*$, then we would have had $\mu_j^l(t) \geq a^*$ for all $t \geq t^*$, since the condition $X_j^l(t) \geq c$ implies that there is at least one complete packet at the originating node of link l. The latter implies that the output rate of stream j from link l is greater than the arrival rate a_j of the stream, which is a contradiction, therefore, t^{**} as above exists.

For any $t_1 > t^{**}$ define

$$t_0(t_1) = \begin{cases} t_1, & \text{if } X_j^l(t_1) \leq c \\ \sup\{t\colon t \leq t_1, X_j^l(t) = c\}, & \text{if } X_j^l(t_1) > c. \end{cases}$$

Note that since $X_j^l(t^{**}) < c$, the time $t_0(t_1)$ is well defined and is greater than or equal to t^{**}. In the following, it is argued that for any t_1, t_2 such that $t_2 > t_1 > t^{**}$

$$\int_{t_1}^{t_2} \mu_j^l(\tau)\, d\tau \leq \int_{t_0(t_1)}^{t_2} \mu_l^{p_j(l)}(\tau)\, d\tau + c - a_j(t_1 - t_0(t_1)). \tag{45}$$

We distinguish the following cases.

A: Assume that $X_j^l(t_1) \leq c$. Then clearly

$$\int_{t_1}^{t_2} \mu_j^l(\tau)\, d\tau \leq \int_{t_1}^{t_2} \mu_l^{p_j(l)}(\tau)\, d\tau + c \tag{46}$$

and (45) follows since $t_0(t_1) = t_1$.

B: Assume that $X_j^l(t_1) > c$. Clearly, we have

$$\int_{t_0(t_1)}^{t_2} \mu_j^l(\tau)\, d\tau \leq \int_{t_0(t_1)}^{t_2} \mu_l^{p_j(l)}(\tau)\, d\tau + c \tag{47}$$

from which we get

$$\int_{t_1}^{t_2} \mu_j^l(\tau)\, d\tau \leq \int_{t_0(t_1)}^{t_2} \mu_l^{p_j(l)}(\tau)\, d\tau + c - \int_{t_0(t_1)}^{t_1} \mu_j^l(\tau)\, d\tau. \tag{48}$$

Notice that by the definition of $t_0(t_1)$, for any t in the time interval $(t_0(t_1), t_1]$, we have $X_j^l(t) \geq c$, therefore, $\mu_j^l(t) \geq a^* > a_j$ for $t \in (t_0(t_1), t_1]$ and from (48), (45) follows.

From (45) and for $t > t^{**}$

$$\begin{aligned}\sup_{t_2>t_1>t}&\left\{\int_{t_1}^{t_2}\mu_j^l(\tau)\,d\tau - a_j(t_2-t_1)\right\}\\ &\le \sup_{t_2>t_1>t}\left\{\int_{t_0(t_1)}^{t_2}\mu_l^{p_j(l)}(\tau)\,d\tau + c\right.\\ &\qquad \left.- a_j(t_2-t_0(t_1))\right\}\\ &\le \sup_{t_B>t_A>\min\{t,t_0(t)\}}\left\{\int_{t_A}^{t_B}\mu_l^{p_j(l)}(\tau)\,d\tau + c\right.\\ &\qquad \left.- a_j(t_B-t_A)\right\}\end{aligned} \tag{49}$$

where the last inequality follows from the fact that if $t_1 > t > t^{**}$, then $t_0(t_1) \ge t_0(t)$. By taking limits in (49), we have

$$\begin{aligned}\lim_{t\to\infty}\sup_{t_2>t_1>t}&\left\{\int_{t_1}^{t_2}\mu_j^l(\tau)\,d\tau - a_j(t_2-t_1)\right\}\\ &\le \lim_{t\to\infty}\sup_{t_B>t_A>\min\{t,t_0(t)\}}\left\{\int_{t_A}^{t_B}\mu_l^{p_j(l)}(\tau)\,d\tau + c\right.\\ &\qquad \left.- a_j(t_B-t_A)\right\}.\end{aligned} \tag{50}$$

Note that $t_0(t)$ is nondecreasing with t by definition. Furthermore, for any τ in the time interval $(t_0(t), t]$, we have $X_j^l(\tau) \ge c$, therefore, $\mu_j^l(\tau) \ge a^* > a_j$ for $\tau \in (t_0(t), t]$. Because of that, it should hold

$$\lim_{t\to\infty} t_0(t) = \infty \tag{51}$$

otherwise, the output rate of stream j at link l will be larger than a_j.

From the fact that $\mu_l^{p_j(l)}$ has eventual burstiness less than or equal to $\hat{b}_j$ and (50) and (51), the theorem follows. ◇

Theorems 1 and 3 imply a bound on the backlog of a traffic stream in the originating node of a link that holds after the initial transient phenomenon and the convergence of the capacity fractions. This is stated in the following theorem.

Theorem 4: If the conditions C1–2 hold, (41) holds and if the stream $\mu_j^{p_j(l)}(t)$ has eventual burstiness less than or equal to $\hat{b}_j$, then the backlog of stream j traffic in front of link l satisfies the condition

$$\limsup_{t\to\infty} X_j^l(t) \le \hat{b}_j + c. \tag{52}$$

Proof: By contradiction, assume that

$$\limsup_{t\to\infty} X_j^l(t) = \tilde{b} > \hat{b}_j + c \tag{53}$$

therefore, for some b^* such that $\tilde{b} > b^* > \hat{b}_j + c$, there exists a sequence $t_n, n = 1, \cdots$ such that

$$\lim_{n\to\infty} t_n = \infty \tag{54}$$

and

$$X_j^l(t_n) \ge b^*. \tag{55}$$

Arguing similarly as in the proof of Lemma 2, we can show that the backlog $X_j^l(t)$ will be less than c infinitely often, or in other words, that there exists a sequence $t'_n, n = 1, 2, \cdots$ such that

$$\lim_{n\to\infty} t'_n = \infty \tag{56}$$

and

$$X_j^l(t'_n) < c. \tag{57}$$

From Theorem 2, (41), and the fact that the stream $\mu_j^{p_j(l)}(t)$ has eventual burstiness less than or equal to $\hat{b}_j$, we understand that there exist ϵ_1, ϵ_2 such that $\epsilon_1 > 0$ and $b^* - \hat{b}_j - c > \epsilon_2 > 0$ and t^* such that

$$f_j^l(t) \ge \frac{a_j}{C_l} + \epsilon_1, \qquad t \ge t^* \tag{58}$$

and

$$\sup_{t_2\ge t_1\ge t^*}\left\{\int_{t_1}^{t_2} a_j(\tau)\,d\tau - a_j(t_2-t_1)\right\} \le \hat{b}_j + \epsilon_2. \tag{59}$$

Select t'_n and t_k such that $t'_n > t^*, t_k > t'_n$ and define

$$\hat{t} = \sup\{t \colon X_j^l(t) < c, t \le t_k\}.$$

Notice that since $t^* < t'_n < t_k$, the time $\hat{t}$ is well defined. Clearly, we have

$$X_j^l(t_k) - X_j^l(\hat{t}) = \int_{\hat{t}}^{t_k}\mu_j^{p_j(l)}(\tau)\,d\tau - \int_{\hat{t}}^{t_k}\mu_j^l(\tau)\,d\tau. \tag{60}$$

Notice that since we assumed the arrival streams cannot have impulses, the backlogs as a function of time cannot have jumps and from the definition of $\hat{t}$, we should have

$$X_j^l(\hat{t}) = c. \tag{61}$$

From (61) and the definition of t_k, we have

$$X_j^l(t_k) - X_j^l(\hat{t}) \ge b^* - c \tag{62}$$

and

$$\int_{\hat{t}}^{t_k}\mu_j^l(\tau)\,d\tau \ge (a_j + \epsilon_1 C_l)(t_k - \hat{t}). \tag{63}$$

Also from the definition of t^* and since $\hat{t} \ge t^*$, we have

$$\int_{\hat{t}}^{t_k}\mu_j^{p_j(l)}(\tau)\,d\tau \le a_j(t_k - \hat{t}) + \hat{b}_j + \epsilon_2. \tag{64}$$

From (60) and (62)–(64), we get

$$a_j(t_k - \hat{t}) + \hat{b}_j + \epsilon_2 \ge b^* - c + (a_j + \epsilon_1 C_l)(t_k - \hat{t})$$

which implies

$$\epsilon_2 \ge b^* - \hat{b}_j - c. \tag{65}$$

Equation (65) contradicts the selection of ϵ_2 that was done earlier and the proof is complete. ◇

The quantity $\lim\ \sup_{t\to\infty}\ X_j^i(t)$ can be viewed as the maximum backlog after the transient phenomena and in analogy with the eventual burstiness, it can be called eventual backlog. Theorem 4 implies that if the arrival process of stream j has eventual burstiness $\hat{b}_j$, then the eventual backlog of the traffic from stream j at link $l_j(k)$ is bounded by $\hat{b}_j + kc$. Under the PF policy, the backlogs of stream j are bounded as long as the arrival stream j has bounded burstiness and the average rates of the other streams exist. So it is possible that the stream j has bounded backlogs, while streams with which it shares the same links may have unbounded backlogs. From Theorem 4, we can see that there is a bound on $X_j^l(t)$ that holds for all $t \geq 0$. This bound will depend on the initial backlogs at $t = 0$, the initial values of the capacity fractions and the update time sequences, in addition to the burstiness of stream j. If the rate proportional processor sharing is employed, then the bound depends only on the burstiness of the stream and the initial condition.

VI. Discussion

The bounds on the backlogs of stream j depend only on the burstiness characteristics of that stream. This is happening because under the PF policy the service fractions eventually converge to values larger than the rates of the corresponding streams, given only the existence of the average rates. After the convergence, the backlogs of every stream depend on the traffic characteristics of the stream only.

Note that at no point in the proof of the results did we make the assumption that a traffic stream can pass through each queue at most once. Hence, the results hold for traffic streams that may be fed back to the same queue several times. The latter case never arises of course in the context of a communication network.

There are several issues to be addressed regarding the applicability of the policies proposed here, for the congestion control of a communication network. How fast do the capacity fractions of each stream converge and how does the convergence time compare to the duration of a session? What is the usefulness of the policy when the rates of the sessions are known and the fractions of the capacity to be allocated to the different streams can be set *a priori*? Finally, how can the policy be implemented without the infinitesimal packet length assumption?

The convergence speed of the capacity fractions depends both on the topology of the network as well as the convergence speed of the time average arrival rates. This is an issue to be studied in association with specific network topologies and applications. If the duration of a session is not long enough compared to the duration of the transient period of the policy, then clearly, the capacity fractions should be preset based on the traffic characteristics of the session.

The PF policy relies on the following principle. Allocate the service capacity of a link in proportion to the traffic loads of the sessions as they are viewed from the link perspective. When the cumulative traffic load from the beginning of the operation of the system is considered, it is shown that we have eventual convergence to capacity fractions that guarantees stability. An alternative is to measure the load within a fixed time window. This will make the policy more volatile and, depending on the window length, it may have the effect of a burst level allocation of the capacity. Instead of allocating the total link capacity in this manner, another option is to allocate portions of the capacity equal to the average rates according to a fixed allocation scheme and the rest of the capacity with the dynamic scheme suggested above. Those alternatives are worth investigation regarding their performance.

Finally an important issue is the implementation of the PF policy under store and forward assumptions on the service and the constraint that, at most, one packet can be transmitted at a time at each link. This issue was investigated for processor sharing by Demers *et al.* in [4] and Parekh and Gallager in [9]. The fair queueing discipline was proposed, which is a packetized service discipline and emulates close processor sharing, as was shown analytically by Greenberg and Madras [5] and Parekh and Gallager [9]. The fair queueing technique is not directly applicable to the PF policy, since the capacity fractions vary with time. The problem of emulating an arbitrary time-varying fluid policy by one that provides packetized service was considered in [11]. A scheme similar to fair queueing was shown to work in the case of time-varying policies as well. That scheme may be used for emulating the PF policy and providing packetized service.

Acknowledgment

The author is grateful to the reviewers for their thoughtful comments that helped to improve the presentation of this paper.

References

[1] C. S. Chang, "Stability, queue length and delay, part I: Deterministic queueing networks," Tech. Rep. RC 17708, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1992.

[2] R. Cruz, "A calculus of network delay, part I: Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 114–131, Jan. 1991.

[3] ——, "A calculus of network delay, part II: Network analysis," *IEEE Trans. Inform. Theory*, vol. 37, pp. 132–141, Jan. 1991.

[4] A. Demers, S. Keshav, and S. Shenkar, "Analysis and simulation of a fair queueing algorithm," in *Proc. SIGCOM'89*, 1989, pp. 1–12.

[5] A. C. Greenberg and N. Madras, "How fair is fair queueing?" *J. ACM*, vol. 3, 1992.

[6] E. L. Hahne, "Round-robin scheduling for max-min fairness in data networks," *IEEE J. Select. Areas Commun.*, vol. 9, no. 7, pp. 1024–1039, Sept. 1991.

[7] S. S. Panwar, T. K. Philips, and M. S. Chen, "Golden ratio scheduling for flow control with low buffer requirements," *IEEE Trans. Commun.*, vol. 40, pp. 765–772, Apr. 1992.

[8] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single node case," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 344–357, June 1993.

[9] ——, "A generalized processor sharing approach to flow control in integrated services networks: The multiple node case, "*IEEE/ACM Trans. Networking*, vol. 2, no. 2, pp. 137–150, Apr. 1994.

[10] L. Tassiulas, "Adaptive back pressure congestion control based on local information," *IEEE Trans. Automat. Contr.*, vol. 40, no. 2, pp. 236–250, Feb. 1995.

[11] ——, "Cut-through switching, pipelining and scheduling for network evacuation," submitted, 1995.

[12] O. Yaron and M. Sidi, "Performance and stability of communication networks via robust exponential bounds," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 372–385, June 1993.

Leandros Tassiulas (S'89–M'91) was born in 1965, in Katerini, Greece. He received the Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Greece in 1987, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1989 and 1991, respectively.

From September 1991 to June 1995 he was an Assistant Professor in the Department of Electrical Engineering, Polytechnic University, Brooklyn, NY. Since July 1995 he has been an Assistant Professor in the Department of Electrical Engineering, University of Maryland, College Park. His research interests are in the field of computer and communication networks with emphasis on wireless communications and high-speed network architectures and management, in control and optimization of stochastic systems and in parallel and distributed processing. He is a consultant to the cellular and satellite industry.

Dr. Tassiulas received a National Science Foundation (NSF) Research Initiation Award in 1992 and an NSF Faculty Early Carrer Development Award in 1995. He also coauthored a paper that received the INFOCOM'94 best paper award.