Extremal traffic and bounds for the mean delay of multiplexed regulated traffic streams

F. M. Guillemin , N. Likhanov R. R. Mazumdar, and C. Rosenberg

***Abstract*—In this paper, we present simple performance bounds for multiplexed regulated traffic streams, which are leaky-bucket regulated with peak, mean rate and burst size constraints. We consider independent, heterogeneous streams, which are multiplexed in a common buffer. We derive bounds on the mean delay in the deterministic context and we then obtain a simple stochastic bound, which is exact when the number of sources increases. A byproduct is a characterization of the worst case sources for mean delay, when they are leaky bucket regulated.**

I. Introduction

With the emergence of the need by users for quality of service (QoS), the basic idea of controlling traffic at the network access has played a crucial role over the past few years in the design of broadband integrated networks [1]. Even though such an approach is constraining because of the difficulty encountered by users to declare traffic parameters, it has prevailed in the development of ATM networks but also in the evolution of the Internet, for instance with the IntServ model and more recently with the standardization of MPLS. Traffic generated by users will thus be conforming to some traffic parameters enforced at network access ; these parameters are negotiated in one way or another between the user and the network.

One of the simplest regulation mechanisms is the so-called leaky bucket mechanism, which has gained enormous popularity in ATM networks. This mechanism has been studied in great depth in the context of providing guaranteed QoS in networks. These regulators are often referred to as (σ, ρ) regulators and a very powerful formalism to study worst-case delay bounds called *network calculus* has been developed for such inputs. The systematic approach goes back to the seminal work of Cruz [2], [3], but has been greatly extended in the works of Le Boudec [4] and Chang [5]. The recent monograph of Chang [6] gives an excellent account of the approach.

Network calculus is essentially a deterministic worst-case approach. An advantage of the approach is that it readily leads to a calculus valid for obtaining an end-to-end worst case delay bound knowing the regulation bounds on the individual streams. However, one important drawback is that, being a deterministic approach, it fails to take into account the fact that traffic streams are usually statistically independent and rarely perfectly synchronized, which is what is assumed for computing the envelope of the multiplexed streams. This negates the effect of statistical multiplexing and leads to very conservative and wasteful allocation of resources if statistical QoS guarantees are only required to be met.

Let us illustrate this point by considering a concrete example. Consider N independent regulated traffic streams which are multiplexed into a common buffer which is drained at c bits per sec. Let $A_j(s,t)$ denote the total number of bits emitted by the jth stream in the interval (s,t). The (σ,ρ) constraint entails that $A_j(s,t) \leq \sigma_j + \rho_j(t-s)$. The parameters ρ_i and σ_i denote the regulation bounds of the leaky-bucket and define bounds on the long-term average rate and the instantaneous size of the bursts from the stream. Let $A(s,t) = \sum_{j=1}^{N} A_j(s,t)$ denote the aggregate multiplexed stream. Then it is trivial to note that the regulation bounds on $A(s,t)$ are provided by $\rho = \sum \rho_i$ and $\sigma = \sum \sigma_i$. Assuming that $\rho < c$, it can be readily seen that the worst-case delay bound is σ/c. In the case of N identical streams, this upper bound becomes $N\sigma/c$ where $\sigma_j = \sigma$.

In this paper, we consider the situation, when N statistically independent regulated traffic streams are multiplexed into a common buffer. We assume that each stream is regulated by a dual leaky-bucket, one bucket controlling the peak rate π and the other one the achievable mean rate ρ, defined with the associated bucket size σ. We thus consider each stream specified by the regulation parameters $(\sigma_j, \rho_j, \pi_j)$, $j = 1, \ldots, N$ where the cumulative input $A_j(0,t)$ in the interval [0,t) from stream j satisfies:

$$A_j(0,t) \leq \min\{\pi_j t, \rho_j t + \sigma_j\}.$$

This model allows us to introduce the peak rate, denoted by π_j, as an explicit part of the envelope characterization.

Recently, there has been much effort in studying the statistical effects of multiplexing regulated sources. The reason for this is the aggregation of individual flows, in the DiffServ categories for example. Most of the emphasis has been on trying to characterize the tail distributions of the queueing delay assuming that independent regulated sources enter a common buffer. In [7], using the fact that the sample-paths are bounded, a Hoeffding type argument is given to characterize the multiplexing effects. The authors also try to characterize the worst case source behavior when many streams are multiplexed. In [8], [9], [10] the authors study the worst-case extremal source behavior for obtaining bounds on the tail distribution. This is based on bounding the moment generating function and the use of the Chernoff bound. These are essentially asymptotic in nature and valid far into the tail only.

This work has been supported in part by a grant from France Telecom through the CTI Programme and by a gift from Nortel Networks

Fabrice Guillemin is with France Telecom R&D, 22300 Lannion, France, e-mail: Fabrice.Guillemin@francetelecom.com

Nikolay Likhanov is with the Institute for Problems of Information Transmission, Russian Academy of Sciences, Moscow 101447, Russia, e-mail: likh1@online.ru

Ravi Mazumdar and Catherine Rosenberg are with the School of Electrical and Computer Engineering, Purdue University, W. Lafayette, IN47907-1285, USA, e-mail: {mazum,cath}@ecn.purdue.edu

In network design, especially for dimensioning and bandwidth allocation for best-effort networks, one quantity of interest is the mean delay [1]. The above mentioned approaches, which are essentially asymptotic in nature, are not appropriate in this context as the initial part of the complementary distribution rather than the tail contributes the most significantly to the mean values and the tail asymptotics do not provide this information, and, moreover are technically not valid in the regions of interest. Hence, there is a need for a fresh approach.

The goal of our work is to provide simple, useful results, which can be used for network dimensioning based on mean values of the delay given that the traffic streams are regulated. Replacing the mean delay by the delay bound obtained from network calculus is too conservative and hence wasteful of resources [11]. For more stringent delay requirements, estimates of the delay tail distribution are required. However, in this paper we restrict our attention to the mean values. The principal contribution is to provide tight estimates for the mean delay when only the (ρ, σ, π) envelope is given and yet exploiting the fact that the sources are statistically independent.

The outline of the paper is as follows: In section II, we formulate the problem and outline the various quantities to be calculated. In Section III, we first consider the deterministic case of a single flow and we show that the worst case input mean delay can be described by an ON-OFF type process given that it satisfies the regulation bounds. This result adds to the well known result for bufferless systems due to Doshi [12]. In Section IV, we extend the results to the case of multiple streams. In Section V, we obtain a bound on the stochastic mean delay and show that the bound is tight when the number of sources is large. This is based on the worst case characterization obtained earlier. Section VI concludes the paper with some general observations on extending the result to a more general situation.

II. Problem formulation

Consider N independent flows multiplexed in a single FIFO server queue with server rate c and assume that flow j, $j = 1, \ldots, N$ is constrained by a $(\sigma_j, \rho_j, \pi_j)$ traffic descriptor, where σ_j, π_j and ρ_j are the parameters of the dual leaky-bucket used to regulate the flow; σ_j is the bucket size, π_j is the peak rate, and ρ_j is the achievable mean rate. Throughout our discussion, we assume a fluid queueing model.

The amount $A_j(0,t)$ of data which is offered by stream j over the time interval $[0,t]$ is a stochastic process defined on some reference stochastic basis $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\})$, where $\{\mathcal{F}_t\}$ is the natural filtration generated by the processes $\{A_j(0,t)\}$ for $j = 1, \ldots, N$. We assume that for all j, $\{A_j(0,t)\}$ is a continuous increasing process with stationary increments.

The $(\sigma_j, \rho_j, \pi_j)$ constraint for stream j consists of assuming that for (almost) every trajectory $\omega \in \Omega$, the amount of data which can be generated by this stream over the time interval (s,t), denoted by $A_j(s,t)$ is such that

$$A_j(s,t) \leq \min\{\pi_j(t-s), \sigma_j + \rho_j(t-s)\}. \tag{1}$$

Let r_t^j be the instantaneous arrival rate of stream j, which equal to the right derivative of the process $\{A_j(0,t)\}$. By definition, $0 \leq r_t^j \leq \pi_j$.

Let w_t denote the amount of fluid in the queue at time t, which is also an $\{\mathcal{F}_t\}$-adapted stochastic process defined on the reference stochastic basis $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\})$. The process $\{w_t\}$ satisfied the evolution equation

$$dw_t = (r_t - c)\left(1 - \mathbb{1}_{\{w_t=0, r_t<c\}}\right) dt, \tag{2}$$

where r_t is the instantaneous arrival rate of the superposition of the N streams, defined by

$$r_t = \sum_{j=1}^{N} r_t^j.$$

Throughout this paper, we assume that the input processes $\{A_j(0,t)\}$ are with stationary and ergodic increments and that the load of the queue, defined by

$$\eta \stackrel{def}{=} \frac{1}{c}\sum_{j=1}^{N} \rho_j < 1.$$

Note that for a given trajectory $\omega \in \Omega$, $t \to A_j(0,t)(\omega)$ for all $j = 1, \ldots N$ and $t \to w_t(\omega)$ are functions defined on $\mathbb{R}_+$ and taking values in $\mathbb{R}_+$.

$A(dt)$ can be seen as a stationary random measure on $\mathbb{R}_+$. Let $\{\theta_t\}_{t\geq 0}$ be a measurable flow on Ω which is $\mathbb{P}$-invariant. Let ρ_A be the average intensity of $A(dt)$, i.e., $\rho_A = \mathbb{E}[A(0,1)] = \sum_{j=1}^{N} \rho_j$.

Associated with the random measure A is a Palm measure $\mathbb{P}_A$ (see [13, Chap. 12.2] and [14], [15], [16]), which is defined as follows: for all $\{\mathcal{F}_t\}$-measurable stationary processes $\{Z(t)\}$ (such that $Z(s) = Z(0) \circ \theta_s$)

$$\mathbb{E}\left[\int_0^t Z(s)A(ds)\right] = \rho_A t \mathbb{E}_A\left[Z(0)\right], \tag{3}$$

where $\mathbb{E}_A$ and $\mathbb{E}$ denote the expectations with respect to $\mathbb{P}_A$ and $\mathbb{P}$, respectively.

Under the assumption $\eta < 1$, there exists a stationary regime for $\{w_t\}$ [14], i.e., there is a unique $\{\theta_t\}$-consistent solution of (2), defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

In the following, we are interested in the mean delay of data through the system. Consider a time interval $(t, t+dt)$, the number of bits generated in this time interval is $r_t dt$, where r_t is the instantaneous arrival rate. These bits experience a delay of w_t/c time units, since the server rate is c. Note that if $w_t = 0$, which is possible in a fluid queue even if data enter the system, these bits experience no delay. Over the time interval $[0,t]$, the total amount of delay experience by all the bits generated in this time interval is

$$\frac{1}{c}\int_0^t w_s A(ds)$$

and the mean delay for the total amount of bits generated over this time interval is

$$\frac{1}{cA(0,t)}\int_0^t w_s A(ds).$$

The performance measure of interest is the long-term average of the above quantity, which under ergodicity, corresponds to

the mean delay seen by an arriving bit, i.e.,

$$\overline{\mathcal{D}} \stackrel{def}{=} \lim_{t\to\infty} \frac{1}{cA(0,t)} \int_0^t w_s A(ds) = \frac{1}{c}\mathbb{E}_A[w_0], \quad (4)$$

where $\mathbb{E}_A[.]$ denotes the expectation w.r.t. the Palm probability defined above.

In the following, we are interested in finding upper bounds for the mean delay defined by equation (4). Let $\{T_j\}$ denote the sequence of times at which the busy periods of the queue begin and let τ_j denote the length of the jth busy period, so that $T_j + \tau_j$ is the ending time of the jth busy period. Let $N_t = \sum_k \mathbb{1}_{\{T_k \le t\}}$; $\{N_t\}$ is the (stationary and ergodic) point process counting the number of busy periods. Then $\{N_t\}$ is also $\{\theta_t\}$-consistent.

Noting that the amount of data waiting in the queue up to time time t is equal to $A(T_1, T_1+\tau_1) + \ldots + A(T_{N_t}, (T_{N_t} + \tau_{N_t})\wedge t)$ is less than or equal to $A(0,t)$, since all the bits arriving in a fluid system does not necessarily queue, $\overline{\mathcal{D}}$ is less than or equal to the limit as $t \to \infty$ of the quantity

$$\frac{N_t}{A(T_1, T_1+\tau_1) + \ldots + A(T_{N_t}, (T_{N_t} + \tau_{N_t}) \wedge t)} \times \frac{1}{cN_t} \sum_{j=1}^{N_t} \int_{T_j}^{(T_j+\tau_j)\wedge t} w_s A(ds),$$

which is equal to (see [14])

$$\mathcal{D} \stackrel{def}{=} \frac{1}{\mathbb{E}_N\left[cA(0,\tau)\right]} \mathbb{E}_N\left[\int_0^\tau w_s A(ds)\right]. \quad (5)$$

where $\mathbb{E}_N[.]$ denotes expectation w.r.t. the Palm probability associated with the process $\{N_t\}$ and τ is the length of a busy period in the stationary regime. By taking into account the fact that the volume of information $A(0,\tau)$ served in a busy period with length τ is equal to $c\tau$, $\mathcal{D}$ can be rewritten as

$$\mathcal{D} = \frac{1}{c^2 \mathbb{E}_N[\tau]} \mathbb{E}_N\left[\int_0^\tau w_s A(ds)\right].$$

Using the inequality

$$\overline{\mathcal{D}} \le \mathcal{D}, \quad (6)$$

we now address the question of obtaining a bound on the mean delay and more precisely on the quantity $\mathcal{D}$. For this purpose, we fix a given trajectory $\omega \in \Omega$ and we determine the extremal behavior of the process $\{w_t\}$ under the (σ,ρ,π) constraint, which maximizes the quantity $\mathcal{D}$ as defined by equation (5). This procedure allows us to obtain a bound on the quantity

$$\frac{1}{c^2\tau} \int_0^\tau w_s A(ds)$$

over a busy period of length τ. Then, by using the ergodicity of the workload process $\{w_t\}$, we derive an upper bound for the stationary mean delay. In other words, we perform a sample path analysis of the extremal behavior of the workload, which yields a bound for the stationary mean delay via the ergodicity of the system.

III. Single source case

In this section, we consider the case of a queue fed by a single fluid source satisfying a (σ,ρ,π)-constraint. The goal of this section is to prove the following result.

Theorem 1: The mean delay $\mathcal{D}$ in a queue fed by a single fluid source satisfying a (σ,ρ,π)-constraint and drained at constant rate c such that $\rho < c < \pi$ is bounded as:

$$\mathcal{D} \le \hat{\mathcal{D}} \stackrel{def}{=} \frac{\sigma}{\rho}\left(1 - \sqrt{\frac{\pi(c-\rho)}{c(\pi-\rho)}}\right). \quad (7)$$

In the above lemma, we assume that $\pi > c$ so that the queue can fill up; otherwise, in a fluid model, the queue would always be empty. Moreover, we assume that $\rho < c$ so as to ensure the stability of the system.

To show Theorem 1, we prove a series of technical lemmas. We start the analysis by determining the traffic pattern which maximizes the average delay in a busy period with fixed length τ, defined by

$$d(\tau) = \frac{1}{c^2\tau} \int_0^\tau w_s A(ds). \quad (8)$$

Lemma 1: In the case of a single source, the traffic pattern, which satisfies the (σ,ρ,π) constraint and which maximizes the mean delay $d(\tau)$ in a busy period with fixed length τ is defined as follows:

- If $\tau \le \pi\sigma/(c(\pi-\rho))$, the extremal traffic pattern is composed of a burst at the peak rate π with length $c\tau/\pi$, followed by silent period with length $\tau(1-c/\pi)$.
- If $\tau > \pi\sigma/(c(\pi-\rho))$, the extremal traffic pattern is composed of a burst at the peak rate π with length $\sigma/(\pi-\rho)$, followed by an activity period at rate ρ and with length

$$\frac{c}{\rho}\left(\tau - \frac{\pi\sigma}{c(\pi-\rho)}\right),$$

and followed in turn by a silent period with length

$$\frac{c-\rho}{\rho}\left(\frac{\sigma}{c-\rho} - \tau\right).$$

Moreover, the length τ of the busy period is such that

$$\tau < \tau_{\max} \stackrel{def}{=} \sigma/(c-\rho). \quad (9)$$

Proof: Let us fix a realization $\omega \in \Omega$ of the stochastic process $\{w_t\}$ and let us consider a given busy period starting, say, at time 0 and ending at time τ. What we have to determine is the traffic pattern which maximizes the quantity $d(\tau)$, defined by equation (8). In other words, we have to find a realization of $w = \{w_t\}_{t\in[0,\tau)}$ so that $d(\tau)$ is maximal. w has to satisfy the constraint:

$$w_t \le \min\{(\pi-c)t, \sigma + (\rho-c)t\} \stackrel{def}{=} \hat{w}_t, \quad (10)$$

which corresponds to the (σ,ρ)-constraint. Of course, since the queue must not empty during the busy period, we have $w_t > 0$ for $0 < t < \tau$. Finally, w must be such that $w_0 = w_\tau = 0$.

Since in a busy period $A(ds) = \dot{w}_t + c$, where

$$\dot{w}_t = \frac{dw_t}{dt},$$

the problem under consideration can be formulated as an optimization problem as follows:

$$\max_w J(w) \stackrel{def}{=} \int_0^\tau f(w_t)dt, \tag{11}$$

where

$$f(w, \dot{w}) = w(\dot{w} + c);$$

w has to satisfy conditions (10), and must be such that $w_t > 0$ for $t \in (0, \tau)$ and $w_0 = w_\tau = 0$.

Let $\mathcal{Y}$ denote the set of admissible solutions to the above optimization problem. Since we deal with a fluid system, $\mathcal{Y}$ is defined as

$$\mathcal{Y} = \{w \in \mathcal{C}^1[0,\tau] : w_0 = w_\tau = 0, \text{ and } 0 < w_t \leq \hat{w}_t,\ 0 < t \leq \tau\},$$

where $\mathcal{C}^1[0,\tau]$ is the set of functions which are continuous over $[0,\tau]$ and derivable over $(0,\tau)$.

For $w \in \mathcal{Y}$, it is easily checked that

$$J(w) = c\int_0^\tau w_t dt,$$

and we see that the optimization problem under consideration consists of finding the element w of $\mathcal{Y}$ such that the area swept under the function $t \to w_t$ is maximal.

Let us define on $\mathcal{Y}$ the partial order $\preceq$ as follows:

$$w \preceq v \quad \text{iff} \quad w_t \leq v_t \text{ for all } 0 \leq t \leq \tau.$$

It is easily seen that $J(w) \leq J(v)$ if $w \preceq v$ and then that the functional J is monotonic increasing.

Let w^* be the element of $\mathcal{Y}$ defined as follows:

- if $\tau \leq \pi\sigma/(c(\pi-\rho))$,

$$w_t^* = \begin{cases} (\pi - c)t,\ 0 \leq t \leq t_1 \stackrel{def}{=} c\tau/\pi, \\ c(\tau - t),\ \tau_1 \leq t \leq \tau. \end{cases} \tag{12}$$

- if $\tau \geq \pi\sigma/(c(\pi-\rho))$,

$$w_t^* = \begin{cases} (\pi - c)t,\ 0 \leq t \leq \tau_1 \stackrel{def}{=} \sigma/(\pi-\rho), \\ \sigma + (\rho - c)t,\ \tau_1 \leq t \leq \tau_2 \stackrel{def}{=} (c\tau - \sigma)/\rho, \\ \sigma + \rho\tau_2 - ct,\ \tau_2 \leq t \leq \tau, \end{cases} \tag{13}$$

The function $t \to w_t^*$ is illustrated in Figure 1 in the case when $\tau \leq \pi\sigma/(c(\pi-\rho))$ and in Figure 2 in the case $\tau \geq \pi\sigma/(c(\pi-\rho))$.

The element w^* is extremal in $\mathcal{Y}$ in the sense that every $w \in \mathcal{Y}$ is such that $w \preceq w^*$. Indeed, in the case $\tau \leq \pi\sigma/(c(\pi-\rho))$ (resp. $\tau \geq \pi\sigma/(c(\pi-\rho))$), owing to the (σ,ρ,π) constraint, $w_t \leq w_t^*$ for all $t \in [0, t_1]$ (resp. $t \in [0, \tau_2]$). Now, assume that there exists some $t_0 \in [t_1, \tau]$ (resp. $t_0 \in [\tau_2, \tau]$) such that $w_{t_0} > w_t$. Then, from Rolle's theorem, there exists some $t_0' \in [t_0, \tau]$ such that $w_{t_0} = -(\tau - t_0)\dot{w}_{t_0'} > w_{t_0}^* = c(\tau - t_0)$, which implies that $\dot{w}_{t_0'} < -c$. This latter inequality is not possible since the drain rate from the queue cannot exceed c. As a consequence, for every $w \in \mathcal{Y}$, we have $w \preceq w^*$. Since the functional J is increasing, the element w^* is the unique solution to the optimization problem (11).

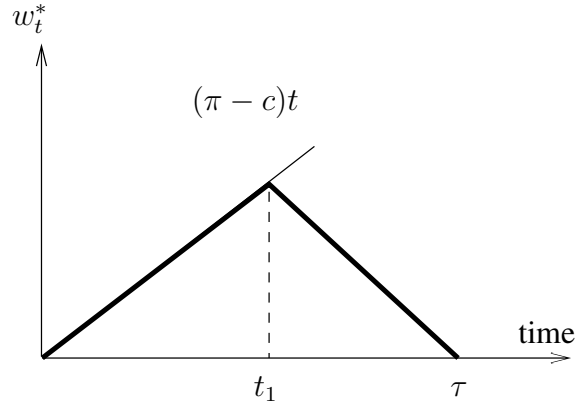

Fig. 1. Graph of the function w^* when $\tau \leq \pi\sigma/(c(\pi-\rho))$ (represented by thick lines).

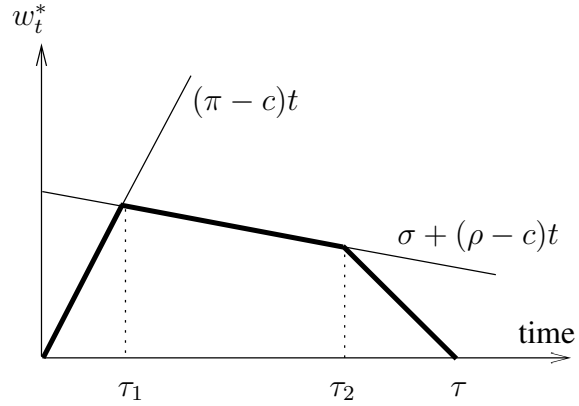

Fig. 2. Graph of the function w^* when $\tau \geq \pi\sigma/(c(\pi-\rho))$ (represented by thick lines).

Now coming back to the input process, when $\tau \leq \pi\sigma/(c(\pi-\rho))$, the input process which maximizes the delay in the busy period with length τ is thus the classical On/Off process; during the On period the arrival rate is equal to the peak rate and the length of the On period is equal to $t_1 = c\tau/\pi$. This is the classical result stating that the optimal control is "bang-bang".

In the case when $\tau \geq \pi\sigma/(c(\pi-\rho))$, the input process, which realizes the optimal trajectory w over a busy period, is as a consequence composed of a burst at the peak rate π and with duration τ_1, followed by an activity period at rate ρ of length $\tau_2 - \tau_1$, and a silent period of length S given by

$$S = \tau - \tau_2 = \frac{c-\rho}{\rho}\left(\frac{\sigma}{c-\rho} - \tau\right). \tag{14}$$

Note that S is positive if and only if $\tau < \sigma/(c-\rho)$. The length of the busy period of a queue with an input process satisfying a (σ,ρ)-constraint is thus necessarily upper bounded by $\sigma/(c-\rho)$. This completes the proof. ■

As a consequence of Lemma 1, we have the following result.

Lemma 2: If the duration of a busy period is τ, the mean delay $d(\tau)$, whatever be the realization of the input process, is

bounded by the quantity $D(\tau)$, where the function D is defined by

$$D(\tau) = \begin{cases} \frac{(\pi-c)\tau}{2\pi}, \ \tau \in [0, \frac{\pi\sigma}{c(\pi-\rho)}], \\ \\ \frac{2\sigma+(\rho-c)\tau}{2\rho} - \frac{\pi\sigma^2}{2\rho c(\pi-\rho)\tau}, \ \tau \in [\frac{\pi\sigma}{c(\pi-\rho)}, \frac{\sigma}{(\rho-c)}]. \end{cases} \tag{15}$$

Proof: Consider a busy period with duration τ. From Lemma 1, we know that the realization of w, which maximizes the mean delay $d(\tau)$ defined by equation (8), is given, when $\tau \leq \pi\sigma/(c(\pi-\rho))$, by the On/Off process composed of bursts at the peak rate π with duration $c\tau/\pi$, followed by a silence period with duration $\tau(1-c/\pi)$. The quantity $d(\tau)$, whatever be the realization of the input process as long as the length of the busy period is $\tau \leq \pi\sigma/(c(\pi-\rho))$, is in this case bounded by:

$$b_1 = \frac{(\pi-c)\tau}{2\pi}.$$

When $\tau > \pi\sigma/(c(\pi-\rho))$, we have to compute the integral $\int_0^{\tau_2} w_s A(ds)$ along the optimal trajectory w illustrated in Figure 2. This integral is equal to

$$\frac{\pi c(\rho-\pi)}{2\rho}\tau_1^2 + \frac{c^2(\pi-\rho)}{\rho}\tau\tau_1 + \frac{(\rho-c)c^2}{2\rho}\tau^2,$$

where $\tau_1 = \sigma/(\pi-\rho)$. After simplification, we have

$$d(\tau) \leq b_2 = \frac{(\rho-c)}{2\rho}\tau + \frac{\sigma}{\rho} - \frac{\pi\sigma^2}{2\rho c(\pi-\rho)\tau}.$$

This completes the proof. ■

The graph of the function D is represented in Figure 3. This function reaches its maximum value at point

$$\tau^* = \sigma\sqrt{\frac{\pi}{(\pi-\rho)(c-\rho)c}} \tag{16}$$

and the maximal value is given by

$$D(\tau^*) = \frac{1}{\rho}(\sigma + (\rho-c)\tau^*). \tag{17}$$

As a consequence, we have for all $\tau \in [0, \sigma/(c-\rho)]$

$$D(\tau) \leq D(\tau^*) = \frac{\sigma}{\rho}\left(1 - \sqrt{\frac{\pi(c-\rho)}{c(\pi-\rho)}}\right). \tag{18}$$

It is easily checked that $\tau^* > \sigma\pi/(c(\pi-\rho))$ and $\tau^* < \tau_{\max}$ since $\pi > c$, where $\tau_{\max}$ is defined by equation (9). Finally, note that

$$D(\tau_{\max}) = \frac{\sigma(\pi-c)}{2c(\pi-\rho)}.$$

The remarkable property of the function D is that its reaches its maximum value at point $\tau^* < \tau_{\max}$. Hence, the maximum value of the upper bound for the mean delay in a busy period is not attained for the maximum length of the busy period (equal to $\tau_{\max}$) but when the length of the busy period is equal to τ^*.

We are now ready to prove Theorem 1.

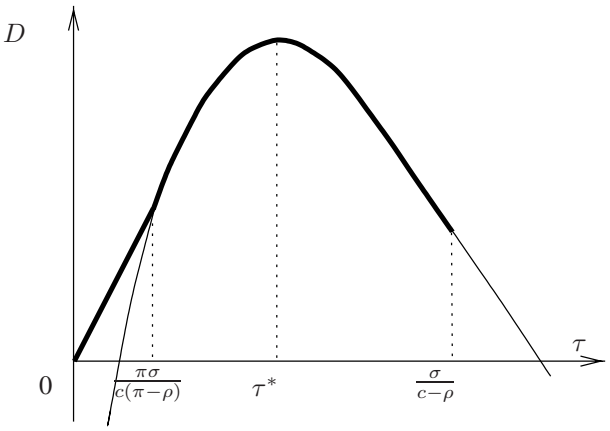

Fig. 3. Graph of function D defined by equation (15) (represented by thick lines).

Proof: [Proof of Theorem 1] We consider a queue fed by a single stationary and ergodic fluid stream satisfying a (σ,ρ,π)-constraint. Let $\{w_t\}$ denote the stochastic processes $\{w_t\}$ describing the amount of fluid in the queue at time t. Since $\rho < c$, the system is ergodic and the mean delay $\overline{\mathcal{D}}$ seen by bits in the system in the stationary regime verifies inequality (6).

From the ergodicity of the system, we have

$$\mathcal{D} = \frac{1}{c^2\mathbb{E}_N[\tau]} \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \int_{T_k}^{T_k+\tau_k} w_s A(ds),$$

From Lemma 2, we know that

$$\frac{1}{c^2}\int_{T_k}^{T_k+\tau_k} w_s A(ds) \leq D(\tau_k)\tau_k \leq D(\tau^*)\tau_k,$$

where T_k and τ_k are the starting time and the length of the kth busy period, respectively and where $D(\tau^*)$ is defined by equation (18). It follows that

$$\mathcal{D} \leq D(\tau^*),$$

and the result follows. This completes the proof. ■

To conclude this section, we can make the following points. So far, we have taken into account constraints on the peak rate and the achievable mean rate. If we relax the constraint on the peak rate, then for a busy period of length τ, the traffic pattern which maximizes the mean delay for the busy period is defined as follows:

- If $\tau \leq \sigma/c$, the traffic pattern y is composed of batches of magnitude $c\tau$ followed by silence periods.
- It $\tau \geq \sigma/c$, the traffic pattern is composed of batches of magnitude σ followed by activity periods at rate ρ and with duration $(c\tau-\sigma)/\rho$, and then by silence periods.

The mean delay in a busy period of length τ is upper bounded by $D_\infty(\tau)$ defined by

$$D_\infty(\tau) = \begin{cases} \tau/2, \ \tau \leq \sigma/c \\ \frac{\sigma}{\rho} + \frac{(\rho-c)}{2\rho}\tau - \frac{\sigma^2}{2c\rho\tau}, \ \frac{\sigma}{c} \leq \tau \leq \frac{\sigma}{c-\rho}. \end{cases}$$

The mean delay in the stationary regime is then upper bounded by

$$\hat{\mathcal{D}}_\infty = \frac{\sigma}{\rho}\left(1 - \sqrt{\frac{(c-\rho)}{c}}\right), \tag{19}$$

which is equal to the upper bound $\hat{\mathcal{D}}$ in Theorem 1 for $\pi = \infty$. The bound $\hat{\mathcal{D}}_\infty$ coincides with the bound reported in [2, Th. 4.7] where it is obtained via direct optimization.

Note that $\hat{\mathcal{D}} < \hat{\mathcal{D}}_\infty$ when $\pi < \infty$. Finally, when $c \gg \rho$, $\hat{\mathcal{D}}_\infty \sim \sigma/2c$ and then, the upper bound is equivalent to the case when data arrive in batches with length σ.

Finally, note that the method used to maximize the mean delay in a busy period with length τ can also be used to maximize the quantity $\int_0^\tau \mathbb{1}_{\{w_t > x\}} dt$ for $x > 0$. We can then obtain an upper bound for the probability distribution of the workload in the stationary regime. This issue has been addressed by Kesidis and Konstantopoulos in [8].

IV. Multiple source case

So far, we have considered the case when there is only one traffic source. Let us now consider the case when N stationary ergodic fluid traffic sources are multiplexed in a FIFO queue drained at constant rate c. The amount of data offered by each source is a stochastic process, which satisfies constraint (1). We assume that the offered load by source j is effectively ρ_j and we denote by $\rho = \sum_j \rho_j$ the total offered traffic. We assume that $\eta \stackrel{def}{=} \rho/c < 1$ so that the system is stable and ergodic.

Lemma 3: The mean delay $\overline{\mathcal{D}}$ in the global queue verifies:

$$\overline{\mathcal{D}} = \frac{\rho_j}{\rho}\overline{\mathcal{D}}_j + \frac{\rho_{\bar{j}}}{\rho}\overline{\mathcal{D}}_{\bar{j}}, \tag{20}$$

where $\overline{\mathcal{D}}_j$ is the mean delay for source j, $\overline{\mathcal{D}}_{\bar{j}}$ is the mean delay for all the other sources, and $\rho_{\bar{j}} = \sum_{k \neq j} \rho_k$.

Proof: Since the system is ergodic, we have

$$\begin{aligned}\overline{\mathcal{D}} &= \lim_{t\to\infty} \frac{1}{A(0,t)} \int_0^t w_s A(ds) \\ &= \lim_{t\to\infty} \frac{1}{A_{\bar{j}}(0,t) + A_j(0,t)} \int_0^t w_s (A_{\bar{j}}(ds) + A_j(ds)),\end{aligned}$$

where $A_j(0,t)$ (resp. $A_{\bar{j}}(0,t)$) is the amount of data generated by source j (resp. all the other sources) over the time interval $(0,t)$. The above equation can be rewritten as

$$\begin{aligned}\overline{\mathcal{D}} = \lim_{t\to\infty} &\frac{A_j(0,t)}{A(0,t)} \frac{1}{A_j(0,t)} \int_0^t w_s A_j(ds) \\ &+ \frac{A_{\bar{j}}(0,t)}{A(0,t)} \frac{1}{A_j(0,t)} \int_0^t w_s A_{\bar{j}}(ds)\end{aligned}$$

Then, by using the definition of the mean delays, we get

$$\overline{\mathcal{D}} = \frac{\rho_j}{\rho}\overline{\mathcal{D}}_j + \frac{\rho_{\bar{j}}}{\rho}\overline{\mathcal{D}}_{\bar{j}}.$$

This completes the proof. ∎

From the above lemma, we easily deduce that

$$\overline{\mathcal{D}} = \frac{1}{\rho} \sum_{j=1}^{N} \rho_j \overline{\mathcal{D}}_j. \tag{21}$$

The total mean delay is thus the weighted sum of the mean delays for the different traffic sources. From the above equation, we see that we can get an upper bound for the global delay if we maximize the mean delay for traffic source j, keeping the other sources unchanged.

Let us fix a realization $\omega \in \Omega$ of the system (i.e., a realization of the different input and queueing processes) and let us consider traffic source j. For the realization ω, $t \to w_t^j(\omega)$ and $t \to w_t^{\bar{j}}(\omega)$ are functions of time t. Let us define a busy period for source j as follows.

Definition 1: A busy period for source j is a time interval over which the function $t \to w_t^j$ is positive, where w_t^j is the amount of data of source j in the queue at time t.

As in the previous section, we have $\overline{\mathcal{D}}_j \leq \mathcal{D}_j$ where $\mathcal{D}_j$ is the mean delay experienced by bits in a stationary busy period, defined by

$$\mathcal{D}_j = \frac{1}{c\mathbb{E}_{N^j}[A(0,\tau^j)]} \mathbb{E}_{N^j}\left[\int_0^{\tau^j} w_s A_j(ds)\right],$$

where τ^j is the length of a source j busy period in the stationary regime and $\mathbb{E}_{N^j}$ is the expectation with respect to the Palm probability $\mathbb{P}_{N^j}$ associated with the point process $\{N_t^j\}$ counting the source j busy periods.

The interaction of source j with the other sources in the queue is seen by source j via the modulation over time of the service rate. Let $c(t)$ denote the service rate of source j traffic at time t. $c(t)$ takes values in $[0,c]$. For the trajectory ω under consideration, $t \to c(t)$ is a given function from $\mathbb{R}_+$ in $[0,c]$.

Lemma 4: For a source j busy period with length τ^j, the traffic pattern of w_t^j which maximizes the mean delay $\mathcal{D}_j$ is defined as follows:

- If $\int_0^{\tau^j} c(t)dt \leq \pi_j\sigma_j/(\pi_j - \rho_j)$, the optimal traffic pattern is composed of burst at the peak rate π_j with duration

$$t_1^j = \frac{1}{\pi_j} \int_0^{\tau^j} c(t)dt.$$

  followed by a silence period with length $\tau^j - t_1^j$.
- If $\int_0^{\tau^j} c(t)dt \geq \pi_j\sigma_j/(\pi_j - \rho_j)$, the optimal traffic pattern is composed of burst at the peak rate π_j followed by an activity period of length $\tau_1^j = \sigma_j/(\pi_j - \rho_j)$, followed by an activity period with duration

$$\frac{1}{\rho_j}\left(\int_0^{\tau^j} c(t)dt - \frac{\pi_j\sigma_j}{\pi_j - \rho_j}\right), \tag{22}$$

  and a silence period with duration

$$S_j = \tau^j + \frac{\sigma_j}{\rho_j} - \frac{1}{\rho_j}\int_0^{\tau^j} c(t)dt.$$

  Moreover, note that the duration τ^j of a source j busy period must be such that

$$\tau^j\left(\frac{1}{\tau^j}\int_0^{\tau^j} c(t)dt - \rho_j\right) \leq \sigma_j.$$

Proof: Let us consider a source j busy period starting, say, at time 0 and ending at time τ^j. As in the proof of Lemma 1, we

have to find the realization of w_t^j, which maximizes the mean delay $\mathcal{D}_j$ given by

$$\mathcal{D}_j = \frac{1}{\tau^j} \int_0^{\tau^j} w_t A_j(dt) \tag{23}$$

In fact, a source j busy period may be composed of several activity periods of source j. Indeed, when source j becomes active and a source j busy period starts, some backlog due to the other sources could be present in the queue and bits of source j will have to wait before being served. The cumulative waiting time could be sufficiently large so as there is an overlap with the next activity period of source j. To study the complete busy period, we have to decompose the busy period into elementary time intervals over which source j is active (i.e., bits from source j arrive at the queue). Over each of these elementary intervals, it is easily checked that the integral in equation (23) is maximal when the input process follows the curve corresponding to the $(\sigma_j, \rho_j, \pi_j)$ constraint, as in the single source case. The optimal $\dot{w}_t^j$ must then be equal to $\pi_j - c(t)$, or $\rho_j - c(t)$ or $-c(t)$, depending on the value of the parameters.

Moreover, one remarkable property is that the time at which the two functions $t \to \int_0^t (\pi_j - c(t))dt$ and $t \to \sigma_j + \int_0^t (\rho_j - c(t))dt$ intersect does not depend on time t and is given by

$$t_1^j = \frac{\sigma_j}{\pi_j - \rho_j}.$$

If the length τ^j of the busy period is such that

$$\int_0^{\tau^j} c(t)dt \leq \pi_j \sigma_j / (\pi_j - \rho_j),$$

the optimal $\dot{w}^j$ is such that $\dot{w}_t^j = \pi_j - c(t)$ over the time interval $[0, \tau_1^j)$ where

$$\tau_1^j = \frac{1}{\pi_j} \int_0^{\tau^j} c(t)dt.$$

and $\dot{w}_t^j = -c(t)$ over the time interval $[\tau_1^j, \tau^j)$.

If $\int_0^{\tau^j} c(t)dt \geq \pi_j \sigma_j / (\pi_j - \rho_j)$, the optimal $\dot{w}_t^j$ is equal to $\pi_j - c(t)$ over the time interval $[0, t_1^j)$, to $\rho_j - c(t)$ over the time interval $[\tau_1^j, \tau_1^j + \tau_2^j]$, and finally equal to $-c(t)$ over the time interval $[\tau_1^j + \tau_2^j, \tau^j)$; τ_2^j is chosen so that

$$\sigma_j + \int_0^{\tau_1^j + \tau_2^j} (\rho_j - c(t))dt - \int_0^{\tau^j} c(t)dt = 0,$$

which implies that

$$\tau_2^j = \frac{1}{\rho_j} \left(\int_0^{\tau^j} c(t)dt - \frac{\pi_j \sigma_j}{\pi_j - \rho_j} \right),$$

The input process which corresponds to this realization of $\dot{w}_t$ is as described in Lemma 4. This completes the proof. ■

A direct consequence of the above lemma is the following result which yields bounds on the mean delay for source j during a busy period of length τ^j.

Lemma 5: If the duration of a source j busy period is τ_j, the mean $d_j(\tau^j)$ delay for source j over this busy period, given by

$$d_j(\tau^j) = \frac{1}{cA(0, \tau^j)} \int_0^{\tau^j} w_s A_j(ds) \tag{24}$$

is bounded, whatever be the realization of the input process, as follows: If $\int_0^{\tau^j} c(t)dt \leq \pi_j \sigma_j / (\pi_j - \rho_j)$,

$$d_j(\tau^j) \leq \frac{1}{2c} A(0, \tau^j) + \frac{\pi_j}{cA(0, \tau^j)} \int_0^{\tau^j} \bar{w_s^j} ds \tag{25}$$

and if $\int_0^{\tau^j} c(t)dt \geq \pi_j \sigma_j / (\pi_j - \rho_j)$, there exists two constants k_1^j and k_2^j such that

$$d_j(\tau^j) = -\frac{1}{2c} \left(1 + \frac{c}{\rho_j}\right) A_j(0, \tau^j) + k_j^1 + \frac{k_2^j}{A_j(0, \tau^j)} + \frac{\pi_j}{cA(0, \tau^j)} \int_0^{\tau^j} \bar{w_s^j} ds. \tag{26}$$

Proof: The proof exploits the evolution of the workload during busy periods and relies on the fact that a superposition of regulated streams is a regulated stream. Details are omitted for the sake of brevity. ■

From the above lemma, we see that

$$d_j(\tau_j) \leq D_j(A(0, \tau_j)) + \frac{\pi_j}{cA(0, \tau^j)} \int_0^{\tau^j} \bar{w_s^j} ds,$$

where the function D_j is defined by

$$D_j(x) = \begin{cases} \frac{x}{2c}, \ x \leq \frac{\pi_j \sigma_j}{\pi_j - \rho_j} \\ -\frac{1}{2c} \left(1 + \frac{c}{\rho_j}\right) x + k_1^j + \frac{k_2^j}{x}, \ x \geq \frac{\pi_j \sigma_j}{\pi_j - \rho_j} \end{cases}$$

The remarkable property of the function D_j is that it is bounded over the interval $[0, \infty)$. It follows that there exists a constant $\mathcal{K}_j$ such that $D_j(x) \leq \mathcal{K}_j$ for all $x \geq 0$. By using the same technique as in the previous section, we obtain the following result.

Corollary 1: The mean delay $\mathcal{D}_j$ for source $\#j$ is upper bounded as

$$\mathcal{D}_j \leq \mathcal{K}_j + \frac{\pi_j}{\mathbb{E}_{N^j}[cA_j(0, \tau^j)]} \mathbb{E}_{N^j} \left[\int_0^{\tau^j} \bar{w_s^j} ds \right]. \tag{27}$$

The asynchronism between the different sources comes through the second term on the right hand side of equation (27), where we have to take the expectation of the workload due to the other sources with respect to the Palm probability measure associated with the busy periods of the source considered. This last term is unfortunately extremely difficult to estimate. This is the main reason why the stochastic bounds developed in the next section are very useful in the context of multiple sources.

V. Stochastic bounds

In the previous sections, we characterized the extremal traffic, which satisfies the (σ, ρ, π) bound, and then found the worst case deterministic delay. We then used this bound to upper-bound the mean delay exploiting the ergodicity of the system.

Another approach is to use the worst case source characterization to obtain a stochastic bound based on the observation that *determinism minimizes waiting times* [14]. Essentially, the result states that the waiting time in a stationary $G/G/1$ queue dominates (in an increasing convex ordering sense) the waiting time in a $G/D/1$ or $D/G/1$ queue with the same mean interarrival and service times [14, Chapter 5.4]. This has been shown for stochastic orders associated with point processes but we conjecture that this result can be extended to the case of fluid queues [15]. To the best of our knowledge the corresponding results have not been fully developed for fluid systems.

Using this result, we can now use the results for fluid queues with On/Off type of inputs, which have been developed in [15], [17], [16]. Note that the mean delay we are interested in is given by $\mathbb{E}_A[w_0]/c$ and from the fluid version of the Little's formula [16, Corollary 4], we have $\mathbb{E}_A[w_0] = \mathbb{E}[w]/\eta$ where $\eta = \rho/c$.

Using the above, we state mean delay result without proof below.

Theorem 2: In a fluid queue with N heterogeneous independent On/Off sources as inputs, with exponentially distributed On periods, under the assumption that $\rho < c$, the mean delay (or waiting time under a FIFO service schedule) is given by:

$$\begin{aligned}\mathcal{D}_b &= \frac{1}{2\rho(c-\rho)}\sum_{i=1}^{N}\frac{1}{m_i}E_{N^i}[F_0^i(L_0^i)-\rho_i L_0^i]^2 \\ &\quad -\sum_{i=1}^{N}\frac{1}{\rho m_i}E_{N^i}\left[\int_0^{L_0^i} t(F_0^i(dt)-\rho_i dt)\right] \quad (28)\end{aligned}$$

where m_i denotes the mean value of a cycle of source i defined as an On period + Off period, $F_0^i(t)$ is the cumulative input on $[0,t]$ for the source i when On under P_{N_i}, L_0^i is the length of an On period of source i, $\rho_i = \mathbb{E}[A_i(0,1)]$ is the average rate of A_i, and $\rho = \sum_{i=1}^N \rho_i$.

Remark 1: The above result assumes that the source On times are stationary, independent r.v's while the silence periods are exponential. Also it is assumed that $F_0(t) > ct$ (i.e., a non-zero workload for the fluid queue can form). This leads to a natural interpretation of the above formula as a Pollaczek-Khinchine type of formula for an $M/G/1$ queue, where $F_i(L_0^i)$ are the"marks", which arrive at the Poisson times corresponding to the end of the Off periods (or beginning of the On periods)

Let us first consider the case of a single source. Let m be the mean of the On+Off periods of the source adjusted such that the mean number of bits is ρ. Assume that the source corresponds to the extremal source in Section III. Then $L_0 = \tau_2 = (c\tau - \sigma)/\rho$ and $F_0(L_0) = c\tau$. Hence, $m = c\mathbb{E}[\tau]/\rho$ and applying the formula above we obtain:

$$\begin{aligned}\mathcal{D}_b &= \frac{\mathbb{E}[F(L_0)-\rho L_0]^2}{2cE[\tau](c-\rho)m} - \frac{(\pi-\rho)\mathbb{E}[\frac{\tau_1^2}{2}]}{cE[\tau]} \\ &= \frac{\sigma^2}{2c(c-\rho)E[\tau]} - \frac{\sigma^2}{2cE[\tau](\pi-\rho)}\end{aligned}$$

Now we take $E[\tau] = \tau^*$, which gives the worst pathwise bound, and we get:

$$\mathcal{D}_b \le \frac{\sigma(\pi-c)}{2c\sqrt{(c-\rho)(\pi-\rho)}}\sqrt{\frac{c}{\pi}}. \quad (29)$$

In the particular case when there is no peak rate constraint (i.e., $\pi \to \infty$), it is easy to see that

$$\mathcal{D}_b(\infty) \le \frac{\sigma}{2\sqrt{c(c-\rho)}}. \quad (30)$$

It can readily be seen that $\mathcal{D} \le \mathcal{D}_b$ so it is indeed a bound. In Figure 4, we have plotted the bounds given by equation (23), (33) and a bound based on a simple batch Poisson model, where batches with size σ arrive according to a Poisson process with rate σ/ρ. It is clear the above estimate provides a much better approximation.

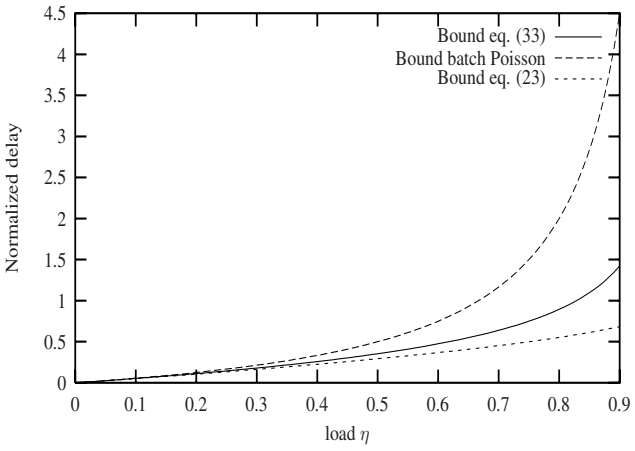

Fig. 4. Accuracy of the stochastic bound $\mathcal{D}_b$.

Let us now address the case of multiplexing N independent, regulated streams. We assume that each source corresponds to the extremal inputs defined in Section III. For this, we apply the formula (28) assuming each source is a worst-case source with mean period $m_i = c\mathbb{E}[\tau_i]/\rho_i$. We take the corresponding mean delay to be the delay bound which we denote by $\mathcal{D}_b^N$

$$\mathcal{D}_b^N = \frac{1}{2(c-\rho)}\sum_{i=1}^{N}\frac{\rho_i\sigma_i^2}{c\rho E[\tau_i]} - \sum_{i=1}^{N}\frac{\rho_i\sigma_i^2}{2\rho cE[\tau_i](\pi_i-\rho_i)} \quad (31)$$

Now taking $\mathbb{E}[\tau_i] = \tau_i^* = \sigma_i\sqrt{\frac{\pi_i}{(\pi_i-\rho_i)(c-\rho_i)c}}$, we obtain

$$\begin{aligned}\mathcal{D}_b^N = \frac{1}{2(c-\rho)}\sum_{i=1}^{N}\frac{\sigma_i\rho_i}{\rho}\sqrt{\frac{(\pi_i-\rho_i)(c-\rho_i)}{\pi_i c}} \\ -\sum_{i=1}^{N}\frac{\sigma_i\rho_i}{2\rho}\sqrt{\frac{c-\rho_i}{c\pi_i(\pi_i-\rho_i)}}\end{aligned}$$

In the case when there is no peak rate constraint, the above formula reduces to:

$$\mathcal{D}_b^N(\infty) = \frac{1}{2(c-\rho)}\sum_{i=1}^{N}\frac{\sigma_i\rho_i}{\rho}\sqrt{\left(1-\frac{\rho_i}{c}\right)} \quad (32)$$

Remark 2: When N is large, $\frac{\rho_i}{c} \approx 0$, in which case the above bound corresponds to the Pollaczek-Khinchine Formula for delay in the $M/G/1$ queue, where the arrivals of type i are

of size σ_i and arrive at a rate σ_i/ρ_i, and the probability, that an arrival is of a type i, is given by ρ_i/ρ. Indeed, when N increases, the load of individual sources decreases and then the bursts of a given source are more and more spread and a Poisson approximation is then justified

Figure 5 shows the simulated mean delay as a function of the number of sources keeping the total load ρ/c fixed. The number of sources was assumed to be N with peak rate 1.01 of which 50% had $\sigma = 20$ while the remaining had $\sigma = 45$. The server speed was assumed to be 1 unit/sec. It is clearly seen that the mean delay approaches the bound given by $\mathcal{D}_b^N(\infty)$ in equation (32) above as is to be expected. It is worth pointing out that the worst case deterministic delay bound cannot be plotted on the same scale in the graph.

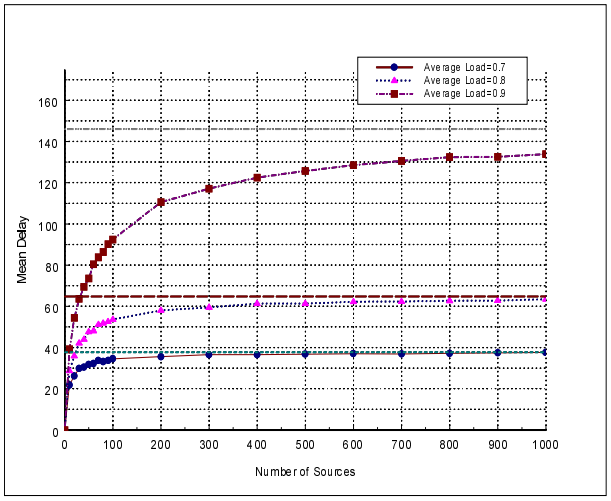

Fig. 5. Accuracy of the stochastic bound $\mathcal{D}_b^N$.

Finally, it is also worth pointing out the gain in considering the stochastic bound over the worst case delay bound when one is interested in dimensioning for mean delays. As mentioned in the introduction, the deterministic worst case bound for N homogeneous sources is $N\sigma/c$, while from above it is roughly $\frac{\sigma}{2(c-\rho)}$, and hence the difference can be considerable when N is large even when the load is high. This shows that there is substantial gain in considering the statistical multiplexing effects even when relatively little statistical information other than the envelopes is available.

VI. Conclusion

Upper bounds for the mean delay experienced by bits of (σ, ρ, π)-regulated sources multiplexed in a single server FIFO queue with a constant service rate c have been derived in this paper. In the case of a single source, we have identified the worst-case source in terms of mean delay. A salient feature of the result is that the maximal value of the mean waiting time is not attained for the classical worst case traffic corresponding to a busy period with length $\tau_{\max} = \sigma/(c-\rho)$, but for a busy period with length τ^* defined by equation (16). The corresponding maximum waiting time is given by equation (7).

In the case of multiple sources, it is also possible to derive deterministic upper bounds for the mean delay. However, these bounds are rather difficult to compute explicitly. Hence, we have developed stochastic upper bounds, which rely on a reasonable conjecture. The stochastic upper bound obtained via this conjecture is given by equation (32). We have showed that the bound is exact when the number of sources increases and can be identified as a simple Pollaczek-Khinchine formula for an $M/G/1$ queue. Moreover, there is substantial gain to be obtained over using the max delay bound when dimensioning buffers for mean delays. A rigorous proof of the conjecture for fluid queues will be addressed in further studies.

Finally, note that a (σ, ρ, π)-constrained source multiplexed with other (σ, ρ, π)-regulated sources should certainly be more regular at the output of the queue than a Poisson batch arrival process. It follows that the stochastic bound conjectured in this paper could be used to develop a network calculus for mean delay through a network of FIFO queues.

Although we have only addressed the mean delay issue in this paper, the results obtained have a very important bearing on obtaining bounds on the delay tail distribution and will be addressed elsewhere.

References

[1] J. Roberts, U. Mocci, and Eds. J. Virtamo, *Methods for the performance evaluation and design of broadband multiservice networks, COST 242 Final Report*, Springer-Verlag, June 1996.

[2] R. L. Cruz, "A calculus for network delay. I. Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 114–131, 1991.

[3] R.L. Cruz, "A calculus for network delay. II. Network analysis," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 132–141, 1991.

[4] J-Y. Le Boudec, "Application of network calculus to guaranteed service networks," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 1087–1096, 1998.

[5] C-S. Chang, "On deterministic traffic regulation and service guarantees: a systematic approach by filtering," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 1097–1110, 1998.

[6] C-S. Chang, *"Performance guarantees in communication networks"*, Springer-Verlag, London, 2000.

[7] L. Massoulié and A. Busson, "Stochastic majorization of aggregates of leaky-bucket constrained traffic streams," Preprint, Microsoft Research, Cambridge, 2000.

[8] G. Kesidis and T. Konstantopoulos, "Extremal traffic and worst-case performance for queues with shaped arrivals," in *"Analysis of communication networks: call centres, traffic and performance (Toronto, ON, 1998)"*, pp. 159–178. Amer. Math. Soc., Providence, RI, 2000.

[9] G. Kesidis and T. Konstantopoulos, "Worst case performance of a buffer with independent shaped arrival processes," *IEEE Comm. Lett.*, vol. 4, no. 1, pp. 26–28, 2000.

[10] K. Kumaran and M. Mandjes, "Multiplexing regulated traffic streams: Design and performance," in *Proc. IEEE INFOCOM 2001*, Achorage, Alaska, 2001.

[11] A. Girard and C. Rosenberg, "Delay and aptimal allocation for max-delay GPS networks," in *Proc. COMCON 8*, Rythmna, Crete, 2001.

[12] B. T. Doshi, "Deterministic rule based traffic descriptors for broadband isdn: Worst case traffic behavior and conection acceptance control," in *Proc. of the ITC 14, Antibes*, pp. 591–600. Elsevier, 1994.

[13] D.J. Daley and D. Vere-Jones, *"An introduction to the theory of point processes"*, Springer Series in Statistics, Springer Verlag, 1988.

[14] F. Baccelli and P. Brémaud, *"Elements of queueing theory"*, vol. 26 of *Applications of Mathematics*, Springer Verlag, New York, 1994.

[15] T. Konstantopoulos and G. Last, "On the dynamics and performance of stochastic fluid systems," *J. Appl. Probab.*, vol. 37, no. 3, pp. 652–667, 2000.

[16] C. Rainer and R. R. Mazumdar, "A note on the conservation law for continuous reflected processes and its application to queues with fluid inputs," *Queueing Systems Theory Appl.*, vol. 28, no. 1-3, pp. 283–291, 1998.

[17] T. Konstantopoulos, M. Zazanis, and G. De Veciana, "Conservation laws and reflection mappings with an application to multiclass mean value analysis for stochastic fluid queues," *Stochastic Process. Appl.*, vol. 65, no. 1, pp. 139–146, 1996.