Network Calculus for Mean Delay Analysis through a Network

Fabrice Guillemin
Orange Labs (France)
fabrice.guillemin@orange.com

Ravi Mazumdar
University of Waterloo (Canada)
mazum@uwaterloo.ca

Catherine Rosenberg
University of Waterloo (Canada)
cath@uwaterloo.ca

Yu Ying
VMware Inc. (US)
yu.ying06@gmail.com

***Abstract*—In this paper, a framework is developed to estimate the mean delay performance of (σ, ρ, π) regulated flows in networks with acyclic routing. We first show that the mean delay performance can be bounded by *on-off* type processes with exponentially distributed *off* periods. We then obtain per-flow bounds on the mean delay. We show that when there is no peak rate constraint, the Pollaczek-Khinchine formula for $M/G/1$ queues provides a tight bound thus establishing the Better-than-Poisson property for such flows. We then consider flows inside a network and show that they can be characterized by a stochastic burstiness parameter and show how the aggregate performance can be bounded from the asymptotic Better-than-Poisson property of regulated flows.**

***Keywords*—Regulated flow; stochastic ordering; mean delay; flow aggregation.**

I. Introduction

Network calculus provides an efficient framework to compute performance bounds for traffic flows traversing a network. This framework has notably been exposed in two standard textbooks [1], [2] published in the early 2000's, following the seminal papers by Cruz [3], [4]. While [2] focuses on deterministic network calculus, [1] formulates network calculus in a probabilistic setting. See also [5] for a more recent introduction to network calculus, where the limitations of this framework are clearly identified.

The basic assumption of network calculus is that traffic flows can be described by means of envelopes, which can be deterministic or stochastic. The most studied envelope is certainly the deterministic one relying on the (σ, ρ, π) parameters, where π is the peak rate, ρ the mean rate and σ the burstiness factor of the flow. With this deterministic envelope, the amount $A(s,t)$ of data that a flow can generate in the time interval (s,t) is such that for all $s \geq 0$

$$A(s,t) \leq \min(\pi(t-s), \rho(t-s)+\sigma). \tag{1}$$

This envelope can be used to determine upper bounds for performance metrics but these bounds are in general conservative.

To obtain tighter bounds, stochastic envelopes have been introduced, notably the Exponentially bounded burstiness (EBB) envelope [6], which is defined as follows: An arrival process $(A(s,t))$ with stationary increments conforms to an EBB envelope with upper rate ρ if for all $s,t > 0$ and all $\sigma \geq 0$

$$\mathbb{P}(A(s,t) \geq \rho(t-s)+\sigma) \leq \kappa e^{-\alpha\sigma} \tag{2}$$

for some positive constants α and κ. The key point is that for EBB flows multiplexed in a buffer, it is possible to compute tight stochastic bounds for the delay of bits in the buffer, see [7] for instance. The bounds are of the type $\mathbb{P}(W > \sigma) \leq e^{-\alpha'\sigma}$. Equation (2) introduces a rather strong assumption on the arrival process for small t while it becomes in practice dummy for large t when flows are ergodic, as $A(s,t)/(t-s) \to \mathbb{E}(A(0,1))$ when $(t-s) \to \infty$, where $\mathbb{E}(A(0,1))$ is the mean offered rate. For small t in contrary, the parameter ρ has to be overestimated when compared to $\mathbb{E}(A(0,1))$ so as Equation (2) applies.

In this paper, we adopt an alternative approach by using stochastic domination in the convex ordering sense. We consider flows that are (σ, ρ, π) regulated, for instance by means of a leaky bucket or a rate limiter as in Equation (1), or can transmit bursts of data more or less periodically as it is the case for TCP connections. Indeed, these connections transmit amounts of data according to the current congestion window size every Round Trip Time (RTT). Even if those amounts can vary at each RTT, they are upper bounded by the maximum window size so that a (σ, ρ, π) envelope can be determined.

In the following, we consider the problem of estimating the mean delay experienced by regulated flows through a network, that was first studied in [8]. We first restrict our attention to the mean delay at network ingress, where it is assumed that FIFO scheduling policy is performed and all traffic flows are independently regulated. For this purpose, we use the fact that regulated flows are dominated by flows with exponential silent periods in the increasing convex ordering sense (see [9] for details). This stochastic ordering property of regulated traffic then leads to a simple and tight upper bound for the mean delay experienced by flows at ingress nodes.

A more challenging issue for estimating the mean delay across a network is to determine what happens in nodes inside the network. The difficulty arises from the fact that traffic flows *inside* the network are usually dependent and their initial statistical properties are altered by multiplexing and scheduling at previous queues. This has led to many studies on how to characterize and compute the queuing performance inside a network in a simple and accurate way [10], [11], [12], [13], [14]. In [11], it is shown that when a large number of homogeneous sources are multiplexed, a single flow at the output of a node essentially keeps its initial statistical characteristics from a large deviations point of view. The

978-0-9883045-5-0/18/$31.00 ©2018 ITC
DOI 10.1109/ITC30.2018.10062

authors in [13] provided a network decomposition framework based on the same idea. With respect to [14], we go further in this paper by considering flows traversing a network and by using the $M/G/1$ queue domination property [15] to obtain bounds on delays through the network.

For this purpose, we consider the concept of stochastic burstiness concept. To obtain tight bounds on delays, it is possible to characterize the stochastic burstiness of single flows and aggregate flows inside the network. We specifically prove that single flows have a stochastic burstiness with constant mean, which is equal to the initial burst size at the network ingress. Its probability distribution spreads (increase in variance) as the number of multiplexing nodes increases, but it converges to its mean if the contribution of this flow is negligibly small at each node, which is the case today with very high speed routers inside networks. This subsequently allows us to use the $M/G/1$ domination property.

The rest of the paper is organized as follows. After recalling a stochastic domination property for regulated flows, we study in Section II the mean delay at a network ingress buffer fed with multiple independent regulated flows. Section III studies the mean delay for single and aggregate flows inside the network. Finally, Section IV concludes the paper.

II. Mean Delay at Network Ingress

A. Stochastic ordering property

For any (σ, ρ, π)-regulated flow, we define an associated exponential *on-off* process $(I^e(s,t))$ as follows.

Definition 1: The input process $(I^e(s,t))$ is a continuous *on-off* process. It has fixed *on* period of length $\sigma/(\pi-\rho)$ at rate π. An on-period is followed by a silent period, which has an exponential distribution with mean σ/ρ.

To state the stochastic ordering property, we first consider the regulated flow $(I_\alpha(s,t))$, which has a periodic profile with a burst at the peak rate with length $\sigma/(\pi-\rho)$, followed by an activity period with length $\alpha > 0$ at rate ρ, followed in turn by a silent period with length σ/ρ. The length of one period is equal to $\beta_\alpha = \sigma/(\pi-\rho) + \alpha + \sigma/\rho$. With the above definitions, the process $(I_\alpha(s,t))$ has the following rate:

$$r_\alpha(s,t) = \begin{cases} \pi & \text{for} \quad 0 \le t \le t_1 = \sigma/(\pi-\rho), \\ \rho & \text{for} \quad t_1 \le t \le t_2 = t_1 + \alpha, \\ 0 & \text{for} \quad t_2 \le t \le t_2 + \sigma/\rho. \end{cases} \tag{3}$$

This profile of $(I_\alpha(s,t))$ follows exactly the envelope defined in Equation (1). It actually produces the maximum amount of data in a time interval $[0, \beta_\alpha]$ among all flows regulated by the same (σ, ρ, π) constraints.

We now assume that the process $(I_\alpha(s,t))$ has a random phase, that is, the starting time of a burst at the peak rate of $(I_\alpha(s,t))$ is uniformly distributed in $[0, \beta_\alpha]$. In this case, we have the following stochastic domination result, which is stated without proof. The actual proof is tedious as numerous scenarios for the quantity $\mathbb{P}\{I_\alpha(s,t) > x\}$ have to be considered. The details can be found in [9], basic elements of increasing convex ordering can be found in [16].

Theorem 1: A (σ, ρ, π)-regulated flow $(I_\alpha(s,t))$ defined by equation (3) is dominated by $(I^e(s,t))$ in the increasing convex ordering sense for all $t \ge 0$, denoted, for short, as $(I_\alpha(s,t)) \le_{icx} (I^e(s,t))$.

We then have the following result for the delay of a bit when the (σ, ρ, π)-regulated flow is injected into a buffer with drain rate c; see [9] for a proof.

Proposition 1: Consider a queue with server rate c and infinite buffer size. Suppose the input is a (σ, ρ, π)-regulated flow with $\rho < c \le \pi$, then the mean delay of an arriving bit chosen at random in the flow entering this queue is upper bounded by

$$\mathcal{D}_b = \frac{\sigma}{2(c-\rho)} \left(1 - \frac{c}{\pi}\right). \tag{4}$$

Simple computations can show that the upper bound given in equation (4) is greater than [8, Equation (1)], which is another mean delay upper bound obtained via sample-path optimization. If the peak rate constraint is now relaxed to $\pi = \infty$, the process $(I_\alpha(s,t))$ turns into a batch arrival of size σ followed by an active period at rate ρ and length α, while $(I^e(s,t))$ becomes a marked Poisson process of mark size σ and rate $\lambda = \rho/\sigma$. Theorem 1, which is in line with the results obtained in [17], then states that regulated flows with batch arrivals are better-than-Poisson in the increasing convex ordering sense. Furthermore, the right hand side of equation (4) is now exactly the Pollaczek-Khinchine formula for an $M/D/1$ queue with the Poisson input $(I^e(s,t))$ and is given by $\mathcal{D}_b(\infty) = \frac{\sigma}{2(c-\rho)}$. We thus re-obtain the result in [17, Corollary 2] for the single flow case.

B. Mean Delay at Server

Let N independent $(\sigma_j, \rho_j, \pi_j)$-regulated traffic sources be multiplexed in a FIFO queue with server rate c. We denote by $\rho = \sum_j \rho_j$ the total mean rate to the queue. We assume $\rho/c < 1$ so that the system is stable. As studied in [8], the asynchronism between different sources makes it very difficult to identify an optimal length α_j for the activity period at rate ρ_j of the jth flow in order to maximize the possible mean delay. In the following, we will use the domination property of exponential flows (given in Definition 1) to obtain an explicit upper bound for the mean delay.

Theorem 2: Consider a fluid FIFO queue with N independent regulated input flows $(I_j(s,t))$, $j = 1, \ldots, N$, the flow $(I_j(s,t))$ being constrained with $(\sigma_j, \rho_j, \pi_j)$ parameters. If $\rho = \sum_{j=1}^N \rho_j < c$ and for all $j = 1, \ldots, N$, $\pi_j \ge c$, then the mean delay at the server is upper bounded by

$$\mathcal{D}_b^N = \sum_{j=1}^N \frac{\rho_j \sigma_j}{2\rho \pi_j} \left(\frac{\pi_j - \rho_j}{c - \rho} - 1 \right). \tag{5}$$

Proof: For each regulated flow $(I_j(s,t))$, we consider the associated exponential flow $(I_j^e(s,t))$ as in Definition 1 . The processes $(I_j^e(s,t))$, $j = 1, \ldots, N$, are mutually independent. Then, we know from Theorem 1 that $(I_j(s,t)) \le_{icx} (I_j^e(s,t))$ for $j = 1, \cdots, N$. Using the independence of the processes $(I_j(s,t))$ and $(I_j^e(s,t))$, we know the superposition

$(I(s,t)) = (\sum_{j=1}^{N} I_j(s,t))$ is dominated by the process $(I^e(s,t)) = (\sum_{j=1}^{N} I_j^e(s,t))$ in the increasing convex ordering sense [16, Chapter 4]. Therefore from [17, Theorem 1], the average workload $\mathbb{E}[w]$ in the queue with regulated inputs is smaller than that with exponential flows, denoted by $\mathbb{E}[w_e]$. Furthermore, from [18], [19], [20],

$$\mathbb{E}[w_e] = \sum_{j=1}^{N} \frac{\rho_j \sigma_j}{2\pi_j} \left(\frac{\pi_j - \rho_j}{c - \rho} - 1 \right).$$

Hence, the mean delay at the queue with regulated inputs is upper bounded by $\mathcal{D}_b^N = \frac{1}{\rho}\mathbb{E}[w_e]$ and the result follows. ∎

Theorem 2 is obtained when the extremal profile of regulated flows is characterized with fixed $(\sigma_j, \rho_j, \pi_j)$ parameters. When flows are characterized with some random parameters, for example, as discussed later in section III, with a random burst size $\tilde{\sigma}_j$, we can adapt results in [18], [19], [20] and Theorem 2 to obtain their mean delay upper bound as

$$\mathcal{D}_r^N = \sum_{j=1}^{N} \frac{\rho_j \mathbb{E}[\tilde{\sigma}_j^2]}{2\rho\pi_j \mathbb{E}[\tilde{\sigma}_j]} \left(\frac{\pi_j - \rho_j}{c - \rho} - 1 \right). \tag{6}$$

This is mainly because increasing convex ordering still holds between regulated flows and exponential flows when burst parameters are independent random variables. This upper bound also implies that the mean and the variance of each $\tilde{\sigma}_j$ suffice for analyzing the mean delay.

When the peak rate goes to infinity, the fluid *on-off* processes converge into batch arrival processes and equation (5) reduces to

$$\mathcal{D}_b^N(\infty) = \frac{\sum_{j=1}^{N} \sigma_j \rho_j}{2\rho(c - \rho)}. \tag{7}$$

This is the Pollaczek-Khinchine formula for delay in an $M/G/1$ queue, where the arrivals of type j have batch size σ_j and arrive at rate ρ_j/σ_j; the probability that a batch arrival is of a type j is given by $\rho_j/(\sigma_j \lambda)$, where $\lambda = \sum_j \rho_j/\sigma_j$. We thus re-obtain results in [17, Corollary 2]. Also note that the bound $\mathcal{D}_b^N$ in Equation (5) is an increasing function of peak rates. This implies that the delay in the $M/G/1$ queue is an upper bound on the delay for regulated flows. This is in line with the observation in [15] that the probability distribution of the workload of regulated flows is asymptotically smaller than that of the $M/G/1$ queue.

Figure 1 shows the mean delay simulated as a function of the number of sources, while keeping the total load ρ/c fixed. This simulation is carried out on a time-driven fluid simulator, which consists of a number of FIFO queues with infinite buffer size and independent regulated *on-off* exogenous input streams. The number of sources was assumed to be N with peak rate $1.01c$ of which 50% had $\sigma = 1$ while the remaining had $\sigma = 10$. The server speed was assumed to be 1 unit/sec. It is clearly seen that the mean delay approaches our bound $\mathcal{D}_b^N$ when N grows. Moreover, in this set of simulations, the numerical values of $\mathcal{D}_b^N(\infty)$ are $9.167, 13.75, 27.5$ for $\rho = 0.7c, 0.8c, 0.9c$ respectively. The quantity $\mathcal{D}_b^N(\infty)$ is an

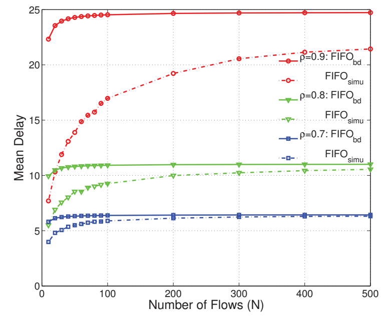


Fig. 1. Accuracy of $\mathcal{D}_b^N$ (5): $c = 1$, $\sigma_1 = 1$,$\sigma_2 = 10$, $\pi_{1,2} = 1.01$, $\rho_{1,2} = \rho/N$.

upper bound for $\mathcal{D}_b^N$ and also bounds the actual mean delay of regulated flows.

The gain from our stochastic bound over the worst case delay bound is substantial. The deterministic worst case bound for N homogeneous sources is $N\sigma/c$ while our result gives roughly $\sigma/(2(c-\rho))$. The difference can be considerable when N is large even when the load is high (e.g., the worst-case delay bound cannot be plotted on the same scale in Figure 1.) This shows that one can take benefit of statistical multiplexing effects to improve the accuracy of mean delay bounds even when relatively little statistical information other than the envelopes is available.

C. Per-class Mean Delay

When the input flows are heterogeneous, i.e. with different $(\sigma_j, \rho_j, \pi_j)$ parameters, the different classes of flows will have different mean delays even with the same FIFO scheduling policy at the queue. Figure 2 illustrates this difference when two classes of flows with different burstiness parameters are considered. The mean delay observed at the server is a weighted average of the delay experienced by different classes of flows, following from the fluid conservation law [18], [19],

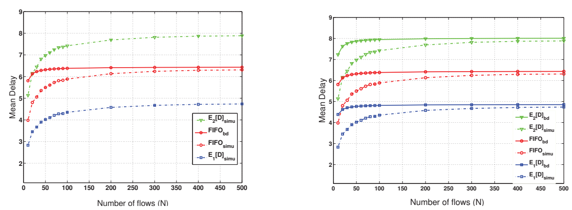
$$\mathcal{D} = \frac{1}{\rho}\mathbb{E}[w] = \sum_{j=1}^{N} \frac{\rho_j}{\rho} \mathbb{E}_{I_j}[\mathcal{D}],$$

where $\mathbb{E}_{I_j}[\mathcal{D}]$ is the Palm distribution associated with the stationary increasing process $(I_j(s,t))$. Recall that the Palm distribution is defined as follows [16]: For an increasing process $(I(s,t))$ and a stationary process $(Z(s,t))$

$$\mathbb{E}_I[Z(0)] = \lim_{t\to\infty} \frac{1}{I(0,t)} \int_0^t Z(0,s) I(0,ds).$$

As also shown in Figure 2(a), the upper bound for the mean delay obtained from Theorem 2 can be much smaller than the actual $\mathbb{E}_{I_2}[\mathcal{D}]$.

From the above discussion, we clearly see for practical purposes that it becomes essential to establish a bound for the mean delay, which is valid per class. The following proposition gives the correct bound per class, as displayed in Figure 2(b); the rather lengthy proof can be found in [9]



(a) Wrong Bound (b) Correct Bound

Fig. 2. Per-class mean delay: $c = 1$, $\sigma_1 = 1$, $\sigma_2 = 10$, $\pi_{1,2} = 1.01$, $\rho_{1,2} = 0.7/N$.

Proposition 2: In a fluid FIFO queue with N types of independent $(\sigma_j, \rho_j, \pi_j)$-regulated flows, under the assumption that $\rho = \sum_{j=1}^{N} \rho_j < c \leq \pi_j$, the mean delay experienced by the jth type of flow is upper bounded by

$$\mathbb{E}_{I_j}[\mathcal{D}] \leq \frac{1}{2c(c-\rho)} \sum_{i=1}^{N} \frac{\sigma_i \rho_i}{\pi_i}(\pi_i - \rho_i - c + \rho) + \frac{\sigma_j}{2c\pi_j}(\pi_j - \rho_j - c + \rho). \quad (8)$$

It is worth noting that Equation (8) shows traffic flows with larger burstiness or higher peak rate experience longer delay as well as flows with lower average rate.

III. Mean Delay inside Networks

A. Stochastic burstiness

To characterize flows traversing buffers, we weaken the notion of burstiness as follows. For any sample path $\omega \in \Omega$ of the stochastic flow $(I(s,t))$, let the stochastic burstiness be defined as in [14] by $\sigma(\omega; s, t) = \max_{s \leq u \leq t}\{I(\omega; u, t) - \rho(t-u)\}$. With this notion of burstiness, when an input process $(I(s,t))$ with mean rate ρ traverses a node with capacity c, the output stream $(\widetilde{I}(s,t))$ is characterized by the traffic descriptor $(\widetilde{\sigma}(s,t), \rho, c)$ so that for any time interval (s,t)

$$\widetilde{I}(s,t) \leq \min\{c(t-s), \widetilde{\sigma}(s,t) + \rho(t-s)\}.$$

Figure 3 illustrates the general scenario for single output flows. The tagged ith input flow $(I_i(s,t))$ traverses k nodes and gives rise to $(\widetilde{I}^i_{k+1}(s,t))$ for node $k+1$. We define the Virtual Arrival Time Process (VATP) $v^i_k(s,t)$ for the ith stream $(\widetilde{I}^i_k(s,t))$ at the kth queue by using the queue occupancy process $(Q_k(s,t))$ as

$$v^i_k(s,t) = \sup_{s \leq t}\left\{ s : s + \frac{Q_k(s)}{c_k} \leq t \right\}.$$

The process $v^i_k(s,t)$ is the arrival time of a bit from source i to the kth queue, which departs at time t. For a FIFO fluid queue, $v^i_k(s,t)$ is a non-decreasing and right-continuous process.

Using the concept of VATP together with the fact that $(\widetilde{I}^i_1(s,t))$ is regulated, the stochastic burstiness $\widetilde{\sigma}^i_{k+1}(s,t)$ for $(\widetilde{I}^i_{k+1}(s,t))$ is

$$\widetilde{\sigma}^i_{k+1}(s,t) = \sigma^i_1 + \sum_{j=1}^{k} \frac{\rho^i_1}{c_j}\left(Q_j(f^k_j(s)) - Q_j(f^k_j(s,t))\right),$$

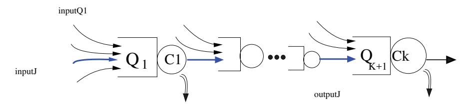


Fig. 3. Single output flow from node k

where $f^k_j(s,t) = v^i_j \circ v^i_{j+1} \circ \cdots \circ v^i_k(s,t)$ is the composition of all the VATPs for the ith flow from node j to node k. The following theorem has been proved in [14].

Theorem 3: In a stable network with work conserving FIFO scheduling at every node with infinite buffer, if all the input flows at the network entrance are regulated with parameters $(\sigma^j, \rho^j, \pi^j)$, then each flow to node k inside the network is regulated with a stochastic burstiness $\widetilde{\sigma}^j_k(s,t)$, which has mean value equal to σ^j and converges to this constant in probability when $\rho^j/c_m \to 0$ for $m = 1, \cdots, k-1$, with c_m being the server capacity of the mth node.

The above result implies that as long as the average arrival rate of the flow is much smaller than the capacity of the server, the burstiness is almost constant through a network. Simulations in [14] have shown that this constant can be a very good approximation, when the network carries a large number of flows, for instance a core network. Hence the mean delay at a node inside the network fed with independent single flows can be estimated by using Theorem 2.

B. Aggregated flow

Assume first all the input flows to a node proceed to the same next node. In this situation, the aggregate flow can be regarded as an *on-off* process. Its *on* periods are the busy periods[1] of the previous queue. Hence, to study the aggregate output flow $(\hat{I}_2(s,t))$ from the first queue, as seen in Figure 4, which consists of a portion of the output flows from Q_1, we can model $(\hat{I}_2(s,t))$ as a periodic *on-off* process. The *on* period is a fraction of a busy period of Q_1 with peak rate c_1, while its *off* period is determined by the silent period of Q_1 and the contribution to the busy period caused by other exogenous input flows to the first queue, which are not aggregated.

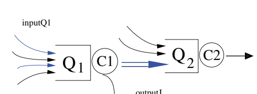


Fig. 4. Aggregate Output Flow of the First Queue

To model $(\hat{I}_2(s,t))$ quantitatively for further analysis, let the busy period of the first queue be denoted by τ_1. The total input flows to Q_1 have an average rate $\rho_1 = \sum_{j=1}^{N_1} \rho^j_1$ and the average rate of the aggregate flow is $\hat{\rho}_2 = \sum_{j=1}^{N_1^A} \rho^j_1$, where N_1^A is the number of flows aggregated at the output of the first queue and injected into the second queue. We then model

[1]In this paper, the busy period of a fluid queue is defined as the time period during which the queue is nonempty in one active cycle.

the length of the *on* period of $(\hat{I}_2(s,t))$ as $Y_1 = \frac{\hat{\rho}_2}{\rho_1}\tau_1$. Since Y_1 now is the maximum time period during which $(\hat{I}_2(s,t))$ produces data at its peak rate c_1, we borrow the previous idea of stochastic burstiness and define a burst size $\hat{\sigma}_2 = (c_1 - \hat{\rho}_2)Y_1$ as the envelope of the data produced by $(\hat{I}_2(s,t))$ in an *on* period. This procedure leads us to define the (stochastic) burst size of the aggregated flow at the output of the first queue as follows.

Definition 2: The stochastic burstiness of an aggregated flow at the output of a queue with service rate c_1 and busy period of length τ_1 is defined as $\hat{\sigma}_2 = (c_1 - \hat{\rho}_2)\frac{\hat{\rho}_2}{\rho_1}\tau_1$, where $\hat{\rho}_2$ is the mean rate of the aggregate and the ρ_1 is the total mean input rate.

The burst size $\hat{\sigma}_2$ is a stationary sequence (from busy period to busy period). With Definition 2 and the fact that the probability of Q_1 being idle is $1 - \rho_1/c_1$, we can easily compute that the average length of the *off* period of $(\hat{I}_2(s,t))$ is $\mathbb{E}[\hat{\sigma}_2]/\hat{\rho}_2$. We thus model the process $(\hat{I}_2(s,t))$ with an *off* period of length $\hat{\sigma}_2/\hat{\rho}_2$.

Proposition 3: The process $(\hat{I}_2(s,t))$ is stochastically regulated with parameters $(\hat{\sigma}_2, \hat{\rho}_2, c_1)$, i.e.,

$$\hat{I}_2(s,t) \leq \min\{c_1(t-s), \hat{\sigma}_2 + \hat{\rho}_2(t-s)\}. \tag{9}$$

Proof: In the time interval $(s,t]$, $\hat{I}_2(s,t) \leq c_1(t-s)$, since the peak rate of $\hat{I}_2(s,t)$ cannot exceed the server rate c_1. Meanwhile, within each *on-off* period, from the definition of $\hat{\sigma}_2$, we know that

$$\hat{I}_2(s,t) - \hat{\rho}_2(t-s) \leq c_1\frac{\hat{\rho}_2}{\rho_1}\tau_1 - \hat{\rho}_2\frac{\hat{\rho}_2}{\rho_1}\tau_1 = \hat{\sigma}_2$$

and the result follows. ■

To obtain the statistical properties of the burst size $\hat{\sigma}_2$ for mean delay analysis, we need to obtain explicit statistical properties of the busy period τ_1 of the first queue, which results from the analysis of a fluid queue with multiple independent *on-off* sources. Explicit results are only available for the average of the workload when these sources have exponentially distributed *off* periods [18], [19], [20]. To estimate the mean delay at the second queue, however, we need the second order statistics of the burst size of the input flows (see equation (6)).

In [15], it is shown that when a large number of regulated flows, say, characterized by the parameters $(\sigma_1^j, \rho_1^j, \pi_1^j)$ for $j = 1, \ldots, N$, enter a FIFO queue with fixed load, the queue length distribution is asymptotically smaller than that of an $M/G/1$ queue, which has Poisson batch arrivals with rate $\lambda_1 = \sum_{j=1}^{N_1} \rho_1^j/\sigma_1^j$ and service time B for the batches defined by $\mathbb{P}\left(B = \frac{\sigma_1^i}{c_1}\right) = \frac{\rho_1^i}{\sigma_1^i \lambda_1}$.

We can thus see that the busy period τ_1 of the queue is asymptotically shorter than the busy period τ_1^p of this $M/G/1$ queue. The Laplace transform $\tau_1^p(s)$ is the solution with modulus less than one to the equation

$$\tau_1^p(s) = B^*(s + \lambda_1 - \lambda_1\tau_1^p(s)), \tag{10}$$

where B^* denotes the Laplace transform of service times B given by $B^*(s) = \frac{1}{\lambda_1}\sum_{j=1}^{N_1} \frac{\rho_1^j}{\sigma_1^j} e^{-s\sigma_1^j/c_1}$.

Using the techniques in [21, Chapter 5.8], we can compute the nth moments of τ_1^p for $n \geq 1$. By approximating τ_1 with τ_1^p, we can then model the statistical moments of our aggregate burstiness $\hat{\sigma}_2$. In particular, the nth moment is given by

$$\mathbb{E}[(\hat{\sigma}_2)^n] = \left(\frac{\hat{\rho}_2(c_1 - \hat{\rho}_2)}{\rho_1}\right)^n \mathbb{E}[(\tau_1^p)^n].$$

The mean value is $\mathbb{E}[\hat{\sigma}_2] = \frac{(c_1 - \hat{\rho}_2)\hat{\rho}_2}{\lambda_1(c_1 - \rho_1)}$ and the second moment

$$\mathbb{E}[(\hat{\sigma}_2)^2] = \frac{(c_1 - \hat{\rho}_2)^2\hat{\rho}_2^2}{\rho_1^2}\frac{c_1\sum_{j=1}^{N_1}\sigma_1^j\rho_1^j}{\lambda_1(c_1 - \rho_1)^3}.$$

In the special case when input flows are homogeneous with common initial burstiness σ_1 and $(\sum_{j=1}^{N_1}(I_1^j(s,t))) = (\hat{I}_2(s,t))$, i.e., $N_1^A = N_1$ and $\hat{\rho}_2 = \rho_1$, we have $\mathbb{E}[\hat{\sigma}_2] = \sigma_1$. This implies that the mean value of the burstiness of the aggregate is equal to the value of the initial burstiness of each input flow. However, the second moment is around $(\sigma_1)^2 c_1/(c_1 - \rho_1) > (\sigma_1)^2$, which indicates that the distribution of $\hat{\sigma}_2$ spreads. Moreover, according to the discussion on the better-than-Poisson asymptotics, this approximation is tight when the number of flows entering the first queue is large.

The above result can be generalized by aggregating a fraction of flows of aggregated flow $(\hat{I}_k^j(s,t))$ with fresh inputs (concentration network). Let the busy period of Q_k be τ_k. We can then define a random burstiness $\hat{\sigma}_{k+1}$ for $(\hat{I}_{k+1}(s,t))$.

Definition 3: Let N_k^A number of the fresh inputs $(I_k^j(s,t))$ and M_k^A number of aggregated inputs from $(\hat{I}_k^j(s,t))$ form the aggregated output $(\hat{I}_{k+1}(s,t))$. The burst size parameter $\hat{\sigma}_{k+1}$ for aggregated flow $(\hat{I}_{k+1}(s,t))$ is defined as $\hat{\sigma}_{k+1} = (c_k - \hat{\rho}_{k+1})\frac{\hat{\rho}_{k+1}}{\rho_k}\tau_k$ with $\hat{\rho}_{k+1} = \sum_{j=1}^{N_k^A}\rho_k^j + \sum_{j=1}^{M_k^A}\hat{\rho}_k^j$ and $\rho_k = \sum_{j=1}^{N_k}\rho_k^j + \sum_{j=1}^{M_k}\hat{\rho}_k^j$.

As in Proposition 3, we can show the following result.

Proposition 4: The process $(\hat{I}_{k+1}(s,t))$ is stochastically regulated according to the parameters $(\hat{\sigma}_{k+1}, \hat{\rho}_{k+1}, c_k)$, i.e.,

$$\hat{I}_{k+1}(s,t) \leq \min\{c_k(t-s), \hat{\sigma}_{k+1} + \hat{\rho}_{k+1}(t-s)\}.$$

Observe that the input flows to the kth queue are regulated flows $(I_k^j(s,t))$, $j = 1, \ldots, N_k$, with fixed bucket sizes and flows $(\hat{I}_k^j(s,t))$, $j = 1, \ldots, M_k$, with random burstiness parameters. The asymptotic better-than-Poisson property for regulated flows with fixed burstiness parameter can be readily extended to this scenario via constructing a Poisson batch arrival process $(X_k^p(s,t))$ with arrival rate

$$\lambda_k = \sum_{j=1}^{N_k}\frac{\rho_1^j}{\sigma_1^j} + \sum_{j=1}^{M_k}\frac{\hat{\rho}_k^j}{\mathbb{E}[\hat{\sigma}_k^j]}.$$

and service time distribution given by

$$B_k = \begin{cases} \sigma_k^i/c_k & \text{w.p. } \frac{\rho_1^i}{\sigma_1^i\lambda_k}, \quad i = 1, \cdots, N_k, \\ \hat{\sigma}_k^i/c_k & \text{w.p. } \frac{\hat{\rho}_k^i}{\lambda_k\mathbb{E}[\hat{\sigma}_k^i]} \quad i = 1, \cdots, M_k. \end{cases}$$

Let τ_k^p denote the busy period of the $M/G/1$ queue when $(X_k^p(s,t))$ enters Q_k. We use the moments of τ_k^p to approximate those of τ_k and model the moments of $\hat{\sigma}_{k+1}$ as follows.

Proposition 5: The moments of the stochastic burstiness at the $(k+1)$th queue are approximated by

$$\mathbb{E}[(\hat{\sigma}_{k+1})^n] \approx \left(\frac{\hat{\rho}_{k+1}(c_k - \hat{\rho}_{k+1})}{\rho_k} \right)^n \mathbb{E}[(\tau_k^p)^n], \ \ \forall n \geq 1.$$

The quantity $\mathbb{E}[(\tau_k^p)^n]$ can be explicitly computed by using the Laplace transform $\tau_k^p(s)$ of the busy period of the $M/G/1$ queue (see Equation (10)), and the Laplace transform of the batch service time distribution $B_k^*(s)$ of the $M/G/1$ queue

$$B_k^*(s) = \frac{1}{\lambda_k} \sum_{j=1}^{N_k} \frac{\rho_k^j}{\sigma_k^j} e^{-s\sigma_k^j/c_k} + \frac{1}{\lambda_k} \sum_{j=1}^{M_k} \frac{\hat{\rho}_k^j (\hat{\sigma}_k^j)^*(s/c_k)}{\mathbb{E}[\hat{\sigma}_k^j]},$$

where $(\hat{\sigma}_k^j)^*$ is the Laplace transform of the random variable $\hat{\sigma}_k^j$. If this random variable can be taken equal to a constant, then $(\hat{\sigma}_k^j)^*(s) = e^{-s(\hat{\sigma}_k^j)^*}$.

To illustrate the above results, we performed simulations for the aggregation as in Figure 4 with parameters given in Table I. The mean delay EW_2^A of aggregate flows at queue 2 is estimated by using Equation (6) and the moments of burstiness factors approximated as above. The average load at each queue is around 0.7.

TABLE I
PARAMETERS FOR FIGURE 4

k	N_k	N_k^A	c_k	σ_k^j	ρ_k^j	π_k^j
1	50:400	$0.2N_1$	10	1	$7/N_1$	11
2	200	—	10	1	6/200	11

It can be observed in Figure 5 that the mean delay estimate gives a reasonably accurate bound, when the number of sources increases. This supports the various approximations made in this paper to approximate the burstiness of flows inside the network and thus their mean delay when traversing the network.

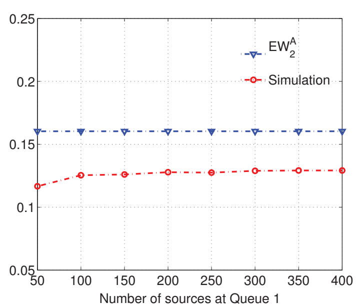


Fig. 5. Mean Delay for Figure 4: Node 2

IV. CONCLUSION

We have developed a framework for the mean delay analysis of acyclic networks consisting of FIFO queues with infinite buffer size and independent input flows which are regulated with (π, σ, ρ) parameters at the entrance to the network. This provides explicit upper bounds for the mean delay of any queue in such networks. Mean delay bounds for queues inside the network are obtained via capturing the statistical properties of traffic inside the network. We proposed the notion of stochastic burstiness and characterized the burstiness behavior of these traffic flows after being altered by the previous scheduling and multiplexing stages. We concluded that when single flows are small, multiplexing does not significantly affect their initial burstiness. The concentration network considered as an application is utmost important for backbone networks concentrating traffic generated by residential customers.

REFERENCES

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

[2] J. L. Boudec and P. Thiran, *Network calulus a theory of deterministic queueing systems for the Internet*, ser. LNCS. Berlin, Germany: Springer Verlag, 2001, vol. 2050.

[3] 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.

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

[5] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Communications Surveys and Tutorial*, vol. 17, no. 1, 2015.

[6] D. Starobinski and M. Sidi, "Stochastically bounded burstiness for communication networks," *IEEE tran. on Information Theory*, vol. 46, no. 1, pp. 206–212, Jan. 2000.

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

[8] F. M. Guillemin, N. Likhanov, R. R. Mazumdar, and C. P. Rosenberg, "Extremal traffic and bounds for the mean delay of multiplexed regulated traffic streams," in *Proceedings of the IEEE INFOCOM 2002*, New York, USA, June 2002.

[9] Y. Ying, "Statistical multiplexing of regulated flows in networks," Ph.D. dissertation, Purdue University, 2006.

[10] L. Massoulié, "Large deviations ordering of point processes in some queueing networks," *Queueing Systems: Theory and Applications*, vol. 28, no. 4, pp. 317–335, 1998.

[11] D. Wischik, "The output of a switch, or, effective bandwidths for networks," *Queueing Systems*, vol. 32, pp. 383–396, 1999.

[12] T. Bonald, A. Proutiere, and J. Roberts, "Statistical performance guarantees for streaming flows using expedited forwarding," in *Proceedings of the IEEE INFOCOM 2001*, Alaska, USA, April 2001.

[13] D. Eun and N.B.Shroff, "Simplification of network analysis in large-bandwidth systems," in *Proc. of IEEE INFOCOM 2003*, San Francisco, CA, March 2003.

[14] Y. Ying, R. Mazumdar, C. Rosenberg, and F. Guillemin, "Burstiness behavior of regualted flows inside networks," in *IFIP Networking 2005*. Waterloo, Canada: Springer, May 2005, pp. 918–929.

[15] Y. Ying, F. Guillemin, R. Mazumdar, and C. Rosenberg, "Buffer overflow asymptotics for multiplexed regulated traffic," *Perform. Eval.*, vol. 65, no. 8, pp. 555–572, 2008.

[16] F. Baccelli and P. Brémaud, *"Elements of queueing theory"*, ser. Applications of Mathematics. Springer Verlag, 2003.

[17] F. M. Guillemin, R. R. Mazumdar, C. P. Rosenberg, and Y. Ying, "A stochastic ordering property for leaky bucket regulated flows in packet networks," *J. of Appl. Prob.*, vol. 44, no. 2, pp. 332 – 348, 2007.

[18] 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 Syst. Theory Appl.*, vol. 28, no. 1-3, pp. 283–291, 1998.

[19] 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.

[20] H. Dupuis and B. Hajek, "A simple formula for mean multiplexing delay for independent regenerative sources," *Queueing Systems*, vol. 16, pp. 195–239, 1994.

[21] L. Kleinrock, *Queueing Systems Vol.1: Theory*. Wiley Interscience, 1975.