

Extremal Traffic and Worst-Case Performance for Queues with Shaped Arrivals

George Kesidis

Dept. of Electrical & Computer Engineering
University of Waterloo
Waterloo, Ontario, N2L 3G1, Canada
`g.kesidis@ece.uwaterloo.ca`

Takis Konstantopoulos

Dept. of Electrical & Computer Engineering
University of Texas at Austin
Austin, TX 78712, USA
`takis@alea.ece.utexas.edu`

Abstract. This paper presents some new results and an overview of the authors' recent work in the area of worst-case performance analysis in communication networks with stationary-ergodic traffic and regulated sample paths. Starting with a single-class network node, the problem of maximizing the buffer overflow probability is considered. Maximization is over a suitable class of stochastic processes. The problem is explicitly solved and the extremal process is identified. Moving on to a two-class queue with jointly stationary arrival processes, the analogous problem is considered. Under some rather natural assumptions the problem is again solved explicitly. The final part of the paper consists of a multi-class queue with the additional constraint that the arrival processes are independent. Bounds on performance measures, such as the tail of the stationary delay distribution, are derived. The problem and results of this paper can be seen as being at the intersection between the effective bandwidths approach and the deterministic network calculus, both of which were introduced in the last decade as tools for evaluating Quality of Service (QoS) in High-Speed Communication Networks. The results can be interpreted in terms of the traffic capacity of rate-regulated channels handling delay-sensitive variable bit rate flows.

1991 *Mathematics Subject Classification.* Primary 90B12, 60G10; Secondary 60K25, 60G17.

Key words and phrases. Stationary processes, communication networks, performance evaluation, optimization, traffic control.

The first author was supported in part by a NSERC of Canada Personal Operating Grant and by Nortel Networks.

The second author was supported in part by NSF Faculty Career Development Award NCR 95-02582 and NSF Grant ANI 99-03495.

1 Introduction

Modern high-speed network standards suggest that a certain degree of deterministic shaping be imposed on traffic. The rationale is that such shaping can help (i) allocate a suitable amount of resources (buffer memory, bandwidth) to a connection to achieve its required Quality of Service (QoS), and (ii) police traffic and assure “fair” access to a shared resource¹. Our models for traffic are right-continuous increasing stochastic processes A with time axis the half line $[0, \infty)$ or the whole real line \mathbb{R} , such that $A(I) < \infty$, a.s., for all finite intervals $I \subseteq \mathbb{R}$. (Note that our notation reflects the fact that we identify increasing functions and Borel measures.) All processes are defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, conveniently endowed with a \mathbf{P} -preserving flow $\{\theta_t\}$. In this paper, all such traffic processes will be stationary in the sense that $A \circ \theta_t(I) = A(I+t)$, for all I and t . By a deterministic shaping constraint we mean a “hard” inequality of the form

$$A(I) \leq g(|I|), \quad (1.1)$$

for all finite intervals I , with $|I|$ being the length of I , and g a suitable deterministic function, e.g., increasing and concave on $[0, \infty)$.

An arbitrary arrival process can be forced to conform to such constraints by delaying its cells using a system of so-called “leaky buckets” in networking practice; see, e.g., [31, 10, 20]. Alternatively, “out of profile” cells or packets may be discarded or marked by using a simple Generic Cell-Rate Algorithm (GCRA) described in Asynchronous Transfer Mode (ATM) standards. Mathematically, a system of leaky-buckets corresponds to a certain non-anticipative causal transformation Ψ that maps an arbitrary traffic process A into one that obeys the constraint above via the rule:

$$\Psi(A)(I) := A(I) \wedge \inf_s [A(I \cap (-\infty, s]) + g(|I \cap (s, \infty)|)].$$

This transformation is, in a sense, optimal. Issues of optimality, realizability, as well as connection to engineering practice can be found in [3]. We mention in passing that to each point in the domain of the convex dual of g there corresponds a leaky bucket.

In what follows we will consider a simple constraint that is actually imposed by communications networking standards:

$$g(t) := \min\{\sigma + \rho t, \pi t\}, \quad t \geq 0, \quad (1.2)$$

where $0 < \rho < \pi$, and $\sigma > 0$. Owing to the fact that $g(t) \rightarrow 0$ as $t \downarrow 0$, our increasing processes are continuous; in addition, they have Lipschitz paths and hence we can talk about their densities, viz., instantaneous transmission rates which exist Lebesgue-a.e. The quantity ρ can be thought of as the mean arrival rate (or an upper bound on the mean arrival rate). Similarly, π can be thought of as the peak rate.

Suppose one or more traffic sources generate constrained traffic of the type above. The role of the network manager is to ensure that

$$\text{Prob}(\text{packet delay} \geq \delta) < \varepsilon. \quad (1.3)$$

This delay constraint can be part of a traffic contract as in, e.g., the case of ATM Variable Bit Rate (VBR) service standards. Alternatively, it can be a general objective of the network, as may be the case in an Integrated or Differentiated Services Internet.

¹See, for instance, www.ietf.org and www.atmforum.com

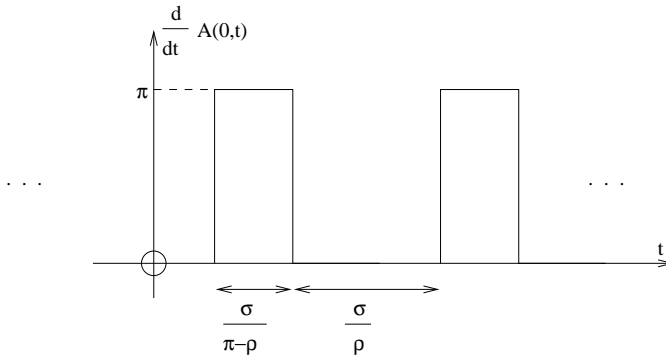


Figure 1 The on-off arrival process is extremal only for a bufferless resource

We pose the following “channel characterization problem”: what is the profile of the worst-case traffic a source can transmit into the network subject to shaping and “quality of service” (QoS) constraints such as (1.3)? For connection admission control (CAC), such “extremal” traffic profiles could be used, instead of other statistical models or actual traces, to provision network resources. Also, extremal traffic can be used to conservatively evaluate performance of network devices. Performance bounds for queues with deterministically shaped arrivals are discussed in [9, 10, 12, 15, 19, 25, 27, 28].

This survey paper is organized as follows: In Section 2, a queue with a single arrival process is considered. The problem is formally posed and the solution, borrowed from [19], is presented. Section 3 presents some new results on a queue with multiple, possibly dependent, arrival processes and a similar problem is solved. The case of a queue with multiple but independent arrival processes is considered in Section 4. In this section, bounds based on moment generating functions are derived. Finally, extensions are discussed in Section 5.

2 The single-source node

The issue of worst-case traffic under (σ, ρ) and π -peak-rate constraints was considered by Doshi [12]. For a bufferless bandwidth resource, Doshi proved that the “extremal” on-off (two rate) arrival process depicted in Figure 1 is “worst-case”, in the sense that it maximizes the proportion of traffic lost. Here, by “worst-case” traffic we mean one that maximizes (1.3)—in a sense to be made precise below. The process is extremal in the sense that the (σ, ρ) constraint is attained at the end of each “burst” (of duration $\sigma/(\pi - \rho)$). However, it was also shown in [12] (by a counter example) that this on-off source is *not* worst-case for a buffered resource.

2.1 Setting things up. Consider then a buffered node with transmission rate $c > \rho$, fed by a stationary-ergodic arrival process A . Let $Q(t)$ be the content of the buffer at time t . Assume that A satisfies (1.1) with g as in (1.2). It can then be seen that the buffer need not be infinite to be lossless. The process Q obeys the dynamics

$$Q(t) = \sup_{0 \leq s \leq t} \{A(s, t] - c(t - s)\} \vee \{Q(0) + A(0, t] - ct\}, \quad (2.1)$$

for all $t \geq 0$. In other words, Q should be the result of the Skorokhod reflection of $\{Q(0) + A(0, t] - ct, t \geq 0\}$. Under the assumption $\rho < c$, there is a unique

Q defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ which is stationary under the flow θ_t , viz., $Q(t) \circ \theta_s = Q(t+s)$ for all t and s .

The usual trick for the construction of a solution to (2.1) is by first extending A to the whole real line (which is possible due to stationarity) and then defining Q , also on the whole real line, by

$$Q(t) := \sup_{-\infty < s \leq t} \{A(s, t] - c(t-s)\}, \quad t \in \mathbb{R}. \quad (2.2)$$

That the Q of (2.2) satisfies (2.1) is obvious. Since $A(s, t] \leq \sigma + \rho(t-s)$ for all $s, t \in \mathbb{R}$, and $\rho < c$, we have that $Q(t) \leq \sigma(\pi - c)/(\pi - \rho) < \infty$ for all t . The same inequality shows that $\lim_{t \rightarrow \infty} \{A(s, t] - c(t-s)\} = -\infty$. This is responsible for the uniqueness of Q . For details see [23]. Having constructed a *stationary* process Q all of whose moments clearly exist (because it is bounded—see below for explicit bounds), define the QoS performance measure

$$\varphi(A) := \mathbf{P}(Q(0) \geq b), \quad (2.3)$$

for a given level $b > 0$. This is closely connected to the stationary probability that the delay of a typical arrival exceeds δ ; indeed, this quantity can be shown to be equal to $\alpha_0^{-1} \mathbf{P}(Q(0) \geq \delta c)$, where α_0 is the ratio of the mean arrival rate $\mathbf{E}A(0, 1)$ over the transmission rate c , see, e.g., [23]. We purposely denote the QoS measure (2.3) as a function of A because our goal is to compute

$$\varphi^* := \sup_A \varphi(A),$$

where the supremum is taken over all stationary-ergodic processes A , deterministically constrained by (1.1), (1.2).

Introduce the notation

$$\alpha := \frac{\rho}{c}, \quad \beta := \frac{\pi - c}{\pi - \rho}, \quad (2.4)$$

and the assumptions

(A1) $\rho < c < \pi$,

(A2) $b \leq \beta\sigma$.

Unless these assumptions hold, the problem can be seen to have a trivial solution. For instance, if $\rho \geq c$ then the system is unstable *for some* arrival process A in our class; if $c \geq \pi$ then there is no queue and $\varphi^* = 0$; if $b > \beta\sigma$, then the queue never reaches level b and so $\varphi^* = 0$. This is also a consequence of the Lemma 2.1 below.

2.2 Probabilistic bounds. The technique for computing φ^* and actually showing that the supremum is achieved, calls for the definition of a “virtual” buffer process X , defined just as in (2.2), but with ρ in place of c :

$$X(t) := \sup_{-\infty < s \leq t} \{A(s, t] - \rho(t-s)\}, \quad t \in \mathbb{R}. \quad (2.5)$$

That X is stationary, ergodic and finite is due to analogous reasoning as that for Q . The following was shown in [19]:

Lemma 2.1 *For all $s, t \in \mathbb{R}$, such that $Q(u) > 0$ for all $u \in (s, t)$, we have*

$$Q(t) - Q(s) \leq \beta(X(t) - X(s)).$$

If, in addition, $Q(t) \geq Q(s)$, we have

$$\frac{Q(t) - Q(s)}{\pi - c} \leq t - s \leq \frac{X(t) - X(s)}{c - \rho}.$$

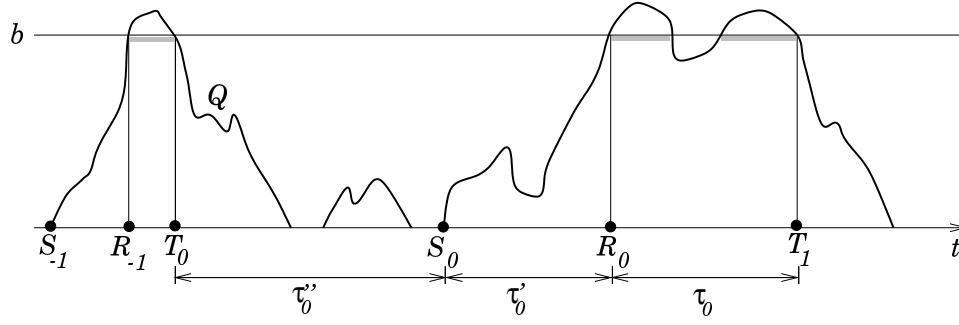


Figure 2 A typical sample path of the Q process

Proof Consider $s < t$ and assume $Q(u) > 0$ for all $u \in (s, t)$. Since $\rho > c$, $X(u)$ is also positive on the same interval. Thus

$$Q(t) - Q(s) = A(s, t) - c(t - s) \quad (2.6)$$

$$\leq (\pi - c)(t - s) \quad (2.7)$$

$$X(t) - X(s) = A(s, t) - \rho(t - s). \quad (2.8)$$

Using (2.6) and (2.8) we have

$$X(t) - X(s) = Q(t) - Q(s) + (c - \rho)(t - s), \quad (2.9)$$

and then, using inequality (2.7), we have

$$X(t) - X(s) \geq Q(t) - Q(s) + \frac{c - \rho}{\pi - c}(Q(t) - Q(s)) = \beta^{-1}(Q(t) - Q(s)),$$

proving the first assertion. The second assertion is trivial: just omit the term $Q(t) - Q(s)$ from (2.9) whenever this is non-negative. \square

In particular, this lemma implies

$$Q(t) \leq \beta X(t),$$

for all t . On noting that

$$X(t) \leq \sigma,$$

for all t (this is a consequence of the constraints (1.2)), it follows then that assumption **(A2)** is essential; if it does not hold, then Q never reaches level b .

The next theorem states an explicit upper bound on $\varphi(A)$ for all allowable processes A :

Theorem 2.2

$$\mathbf{P}(Q(0) \geq b) \leq \frac{\sigma - \beta^{-1}b}{\alpha^{-1}\sigma - b}.$$

Proof [19] The method of proof is based on breaking the process Q , the result of an arbitrary feasible A , into “cycles” as depicted in Figure 2. Consider the down-crossings times of level b . These times form a point process with positive rate, due to **(A2)**. A down-crossing time is called *special* if it is followed by a visit to level 0 before the next down-crossing time. Let T_n be the n -th special down-crossing

time after time 0, $n = 1, 2, \dots$. Also enumerate the special down-crossing times backwards, before 0. The n -th cycle is the piece of the process Q

$$\{Q(t), \quad T_n \leq t < T_{n+1}\}.$$

We also need to introduce the points

$$\begin{aligned} R_n &:= \inf\{t > T_n : Q(t) \geq b\}, \\ S_n &:= \sup\{t < R_n : Q(t) = 0\}, \end{aligned}$$

and so $T_n < S_n < R_n \leq T_{n+1}$, for all $n \in \mathbb{Z}$. Finally, let

$$\tau_n := T_{n+1} - R_n, \quad \tau'_n := R_n - S_n, \quad \tau''_n := S_n - T_n.$$

Next consider \mathbf{P} conditional on $\{T_0 = 0\}$, and call it \mathbf{P}^\sharp . This conditional probability, rigorously defined as a Radon-Nikodým derivative of a certain type, is known as Palm probability with respect to the point process $\{T_n, n \in \mathbb{Z}\}$. Now the following “inversion” formula is the key to obtaining bounds:

$$\varphi(A) = \frac{\mathbf{E}^\sharp \int_{T_0}^{T_1} \mathbf{1}(Q(s) \geq b) ds}{\mathbf{E}^\sharp(T_1 - T_0)}$$

where \mathbf{E}^\sharp is expectation with respect to \mathbf{P}^\sharp . It is easily seen that

$$\varphi(A) \leq \left(1 + \frac{\mathbf{E}^\sharp(\tau'_0 + \tau''_0)}{\mathbf{E}^\sharp(\tau_0)}\right)^{-1}. \quad (2.10)$$

We now bound the terms in the last fraction by repeatedly using Lemma 2.1. Since $Q > 0$ on (S_0, R_0) ,

$$X(R_0) - X(S_0) \geq \beta^{-1}(Q(R_0) - Q(S_0)) = \beta^{-1}b.$$

Since $Q > 0$ on (R_0, T_1) , and $Q(R_0) = Q(T_1) = b$,

$$\begin{aligned} \tau_0 &\leq \frac{X(T_1) - X(R_0)}{c - \rho} = \frac{X(T_1) - X(S_0) + X(S_0) - X(R_0)}{c - \rho} \\ &\leq \frac{X(T_1) - X(S_0) - \beta^{-1}b}{c - \rho}. \end{aligned} \quad (2.11)$$

Since $Q > 0$ on (S_0, R_0) , and $0 = Q(S_0) \leq Q(R_0) = b$,

$$\tau'_0 \geq \frac{Q(R_0) - Q(S_0)}{\pi - c} = \frac{b}{\pi - c}. \quad (2.12)$$

Next, using the inequality

$$X(S_0) - X(T_0) \geq A(T_0, S_0) - \rho(S_0 - T_0) \geq -\rho\tau''_0,$$

which is intuitively obvious, but also formally follows from (2.5), we have

$$\tau''_0 \geq \frac{X(T_0) - X(S_0)}{\rho}. \quad (2.13)$$

Observe (Palm stationarity) that $\Delta := \mathbf{E}^\sharp(X(T_0) - X(S_0)) = \mathbf{E}^\sharp(X(T_1) - X(S_0))$, take \mathbf{E}^\sharp -expectations in (2.11), (2.12), (2.13), and substitute in (2.10) to obtain

$$\varphi(A) \leq \left(1 + \frac{\frac{\Delta}{\rho} + \frac{b}{\pi - c}}{\frac{\Delta - \beta^{-1}b}{c - \rho}}\right)^{-1} = \frac{\Delta - \beta^{-1}b}{\alpha^{-1}\Delta - b}.$$

The fact that the latter expression is increasing in Δ , allows us to substitute Δ by its maximum value σ to obtain the desired bound. \square

2.3 Solving the optimization problem. An arrival process that actually achieves this bound (and hence solves the optimization problem) will now be constructed using a “greedy” approach. To this end, start the system empty and have the source transmit at maximum rate π until the buffer process Q^* reaches level b . This takes an amount of time

$$\tau'_0 = \frac{b}{\pi - c}.$$

At this point, switch the source down to the transmission rate c and keep it at this rate until the slackness in the (σ, ρ) constraint is extinguished, i.e., until X^* reaches level σ . This takes an additional amount of time

$$\tau_0 = \frac{\sigma - \beta^{-1}b}{c - \rho}.$$

During this time, the process Q^* naturally stays at level b without ever exceeding it. After that, switch the source off until the slackness in the (σ, ρ) constraint is maximized, i.e., until X^* reaches 0. This takes an amount of time

$$\tau''_0 = \frac{\sigma}{\rho}.$$

At the end of $\tau'_0 + \tau_0 + \tau''_0$, the process is repeated periodically. Of course, stationarity needs an extra randomization of phase, i.e., in addition to the above arrange so that the origin of time, 0, is uniformly chosen over a cycle. So, we have defined a stationary process $A^* \in \mathcal{A}_{\sigma, \rho, \pi}$ that satisfies the constraints (1.1), (1.2) and achieves the bound of Theorem 2.2 (check by using the formula $\varphi(A^*) = \tau_0 / (\tau'_0 + \tau_0 + \tau''_0)$ together with the expressions above.)

2.4 Delay maximization. By (steady-state) delay we mean the delay of a “typical arrival”. Formally, the probability that the delay exceeds δ is “the fraction of arrivals that see a buffer load larger than $c\delta$ ”. In modern Probability, this is best captured by first defining $D = Q(0)/c$ and then considering the distribution of D under \mathbf{P} conditional on the event that the source is not idle at time 0. This is tantamount to considering the \mathbf{P}_A -expectation of D where \mathbf{P}_A is the Palm probability of P with respect to (the random measure) A . We briefly mentioned above that $\mathbf{P}_A(D > \delta) = \alpha_0^{-1} \mathbf{P}(Q(0) > c\delta)$. Since, for A^* we have

$$\mathbf{E} A^*(0, 1) = \rho$$

we obtain

$$P_{A^*}(D^* \geq \delta) = \frac{\sigma - \beta^{-1}c\delta}{\sigma - \alpha^{-1}c\delta}. \quad (2.14)$$

That this is the worst-case delay bound needs a further proof. This is done in [19]. The technique is similar to the one outlined above, except that one starts with

$$P_A(D \geq \delta) = \frac{\mathbf{E}^\dagger \int_{T_0}^{T_1} \mathbf{1}(Q(s) \geq c\delta) A(ds)}{\mathbf{E}^\dagger A(T_0, T_1]}.$$

Bounds are obtained by breaking the process into cycles. Finally, the worst-case process is as in Figure 3 with $b = c\delta$.

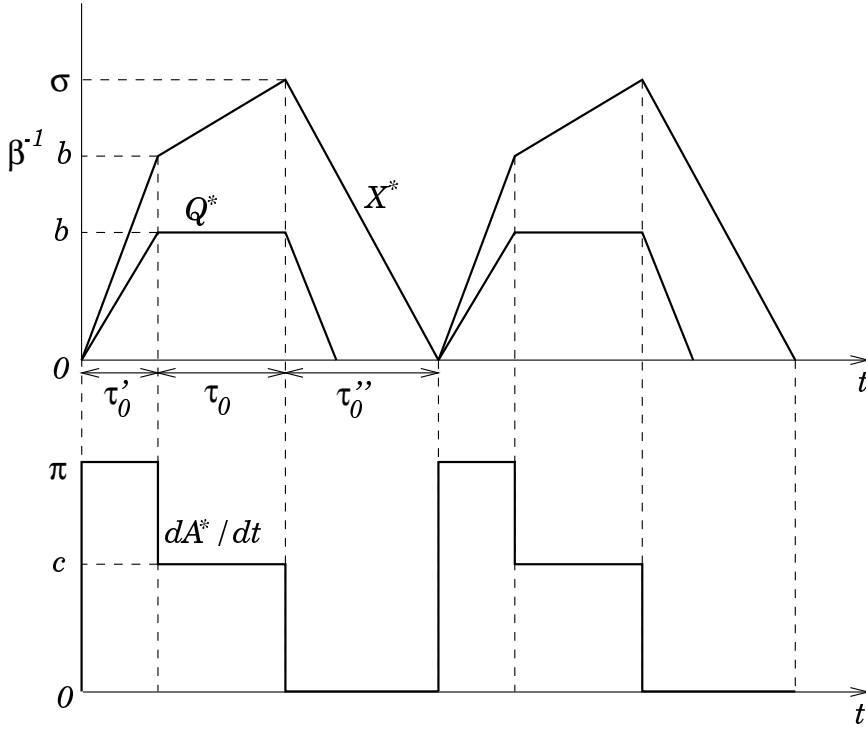


Figure 3 An extremal arrival process for the single-class case

2.5 Interpretations. Suppose a queue is handling shaped VBR traffic with QoS requirement

$$P(D \geq \delta) \leq \varepsilon \quad (2.15)$$

for some positive $\varepsilon < 1$. By (2.14), the required QoS is achieved if the queue's service rate (bandwidth allotment) c solves

$$\varepsilon = \frac{\sigma - \beta^{-1}c\delta}{\sigma - \alpha^{-1}c\delta}$$

where we recall that $\beta = (\pi - c)/(\pi - \rho)$ and $\alpha = \rho/c$. The solution c^* is, of course, a function of the traffic shaping parameters σ, ρ, π and QoS parameters δ, ε . Therefore, one can interpret c^* as being a worst-case *effective bandwidth* of the VBR source.

In this same context, suppose the service rate of the queue is c^* . The extremal arrival process of Figure 3 characterizes the VBR *channel capacity* or maximum throughput (or “goodput”) subject to the (σ, ρ) and $(0, \pi)$ shaping constraints and to the QoS requirement (2.15).

3 A Multiclass Queue

In this section we examine the effect of two sources feeding a buffer and the corresponding worst-case scenario. As before, the transmission rate is constant c . The arrival process is the superposition (aggregation) of two deterministically

constrained arrival processes, A_1, A_2 , jointly stationary and ergodic. A_i is deterministically constrained by

$$g_i(t) := \min\{\sigma_i + \rho_i t, \pi_i t\}, \quad t \geq 0, \quad (3.1)$$

for $i = 1, 2$. We set $A := A_1 + A_2$ for the superposition. The interpretation we have in mind is a 2-class networking node with class 1 being the traffic of interest (under consideration), and class 2 being the “background” traffic which may by itself be the accumulation of several traffic streams.

By a simple additivity property, the aggregate arrival process A is constrained by

$$g(t) = \min\{\sigma_1 + \sigma_2 + (\rho_1 + \rho_2)t, \sigma_1 + (\rho_1 + \pi_2)t, \sigma_2 + (\rho_2 + \pi_1)t, (\pi_1 + \pi_2)t\},$$

where $t \geq 0$. As in the single-class case, we assume that

$$\begin{aligned} 0 < \rho_i < \pi_i, \quad \sigma_i > 0, \quad i = 1, 2, \\ \rho := \rho_1 + \rho_2 < c < \pi := \pi_1 + \pi_2. \end{aligned} \quad (3.2)$$

Let $Q(t)$ be the contents at time t of the shared queue defined as in (2.2). It follows directly that

$$Q(t) \leq Q^{\max} := \max_{t \geq 0} \{g(t) - ct\}.$$

The process $\{Q(t), t \in \mathbb{R}\}$ is stationary and ergodic with deterministically bounded sample paths.

Consider the stationary process Q and break it into cycles, as in the proof of Theorem 2.2. We refer again to Figure 2. Let D_1 be a random variable representing the steady-state class 1 (the class of interest) delay. For a given b such that $b \leq Q^{\max}$, we are interested in maximizing

$$\varphi(A_1, A_2) := \text{Prob}(D_1 \geq bc^{-1})$$

over all A_1, A_2 satisfying the constraints above. Recalling that $\text{Prob}(D_1 \geq bc^{-1})$ is nothing else but $\mathbb{P}_{A_1}(Q(0) \geq b)$, and using a Palm inversion formula, we obtain

$$\varphi(A_1, A_2) = \frac{\mathbb{E}^\dagger \int_{T_0}^{T_1} \mathbf{1}(Q(s) \geq b) A_1(ds)}{\mathbb{E}^\dagger A_1(T_0, T_1)} = \left(1 + \frac{\mathbb{E}^\dagger \int_{T_0}^{T_1} \mathbf{1}(Q(s) < b) A_1(ds)}{\mathbb{E}^\dagger \int_{T_0}^{T_1} \mathbf{1}(Q(s) \geq b) A_1(ds)} \right)^{-1} \quad (3.3)$$

$$\leq \left(1 + \frac{\mathbb{E}^\dagger A_1(S_0, R_0)}{\mathbb{E}^\dagger A_1(R_0, T_1)} \right)^{-1} = \frac{\mathbb{E}^\dagger A_1(R_0, T_1)}{\mathbb{E}^\dagger A_1(S_0, T_1)}. \quad (3.4)$$

We proceed in lower-bounding $A_1(S_0, R_0)$ and upper-bounding $A_1(R_0, T_1)$, but we first introduce some further assumptions to simplify the subsequent calculations.

(A3) $\pi_2 > c$

(A4) $b > \frac{\pi_2 - c}{\pi_2 - \rho_2} \sigma_2$

It is easy to check that the allowable range of values of b in **(A4)** is nonempty can be deduced from the inequality

$$\frac{\pi_2 - c}{\pi_2 - \rho_2} \sigma_2 < Q^{\max}.$$

This inequality follows from the basic inequalities (3.2) and the definition of Q^{\max} . Assumption **(A3)** may be reasonable when the volume of background traffic is large compared to the class-1 traffic. Assumption **(A4)** is intended to prevent the

situation where background traffic alone can reach level b ; if this can happen, then 1 is a tight upper bound on (3.3). Assumption (A4) will be discussed further at the end of this section.

3.1 Lower bound on $A_1(S_0, R_0)$. We refer again to Figure 2. Since $Q(t) > 0$ for all $t \in (S_0, R_0]$ we have

$$A_1(S_0, R_0) = b + c\tau'_0 - A_2(S_0, R_0)$$

Using the constraint on A_2 , i.e., $A_2(S_0, R_0) \leq \sigma_2 + \rho_2(R_0 - S_0)$, we further have

$$A_1(S_0, R_0) \geq b - \sigma_2 + (c - \rho_2)\tau'_0. \quad (3.5)$$

Alternatively, by the constraint $A_2(S_0, R_0) \leq \pi_2(R_0 - S_0)$,

$$A_1(S_0, R_0) \geq b - (\pi_2 - c)\tau'_0. \quad (3.6)$$

Assumption (A3) and (3.6) imply

$$\tau'_0 \geq \frac{b - A_1(S_0, R_0)}{\pi_2 - c},$$

which, with (3.5), implies

$$A_1(S_0, R_0) \geq \left(\frac{\pi_2 - \rho_2}{\pi_2 - c} b - \sigma_2 \right) \frac{\pi_2 - c}{\pi_2 - \rho_2} = b - \frac{\pi_2 - c}{\pi_2 - \rho_2} \sigma_2 > 0 \quad (3.7)$$

where the last inequality is Assumption (A4).

3.2 Upper bound on $A_1(R_0, T_1)$. We will first find an upper bound on $A_1(S_0, T_1)$. From the constraints we have

$$A_i(S_0, T_1) \leq \sigma_i + \rho_i(T_1 - S_0), \quad i = 1, 2. \quad (3.8)$$

Now, $Q(S_0) = 0$, $Q(T_1) = b$ and $Q(t) > 0$ for all $t \in (S_0, T_1]$. Thus,

$$b = Q(T_1) - Q(S_0) = A_1(S_0, T_1) + A_2(S_0, T_1) - c(T_1 - S_0).$$

This equality and (3.8) imply

$$A_2(S_0, T_1) = b + c(T_1 - S_0) - A_1(S_0, T_1) \leq \sigma_2 + \rho_2(T_1 - S_0).$$

The latter inequality directly yields

$$T_1 - S_0 \leq \frac{\sigma_2 - b + A_1(S_0, T_1)}{c - \rho_2}. \quad (3.9)$$

Substituting (3.9) into (3.8) we have

$$A_1(S_0, T_1) \leq \sigma_1 + \rho_1 \frac{\sigma_2 - b + A_1(S_0, T_1)}{c - \rho_2},$$

whence

$$A_1(S_0, T_1) \leq \frac{c - \rho_2}{c - \rho} \sigma_1 + \frac{\rho_1}{c - \rho} (\sigma_2 - b).$$

Finally, by this inequality and (3.7),

$$\begin{aligned} A_1(R_0, T_1) &= A_1(S_0, T_1) - A_1(S_0, R_0) \\ &\leq \frac{c - \rho_2}{c - \rho} (\sigma_1 - b) + \left(\frac{\rho_1}{c - \rho} + \frac{\pi_2 - c}{\pi_2 - \rho_2} \right) \sigma_2. \end{aligned} \quad (3.10)$$

3.3 Worst-case delay and extremal traffic processes. Combining the Palm inversion formula expression (3.4) for $\varphi(A_1, A_2)$, the lower bound (3.7) and the upper bound (3.10), we have therefore proved

Theorem 3.1 *For the two-class system introduced above, under the basic assumptions (3.2) and the additional assumptions (A3) and (A4), the stationary probability that the delay of class-1 traffic exceeds bc^{-1} has an explicit upper bound:*

$$\varphi(A_1, A_2) \leq \frac{\frac{c-\rho_2}{c-\rho}(\sigma_1 - b) + \left(\frac{\rho_1}{c-\rho} + \frac{\pi_2 - c}{\pi_2 - \rho_2}\right)\sigma_2}{\frac{c-\rho_2}{c-\rho}\sigma_1 + \frac{\rho_1}{c-\rho}(\sigma_2 - b)} =: \varphi^*, \quad (3.11)$$

for all jointly stationary, deterministically constrained, as in (3.1), processes A_1, A_2 .

We now pass on to the question of finding whether the bound (3.11) is achievable. Unfortunately, we can only do that under the additional assumption

(A5) $\pi_1 + \rho_2 > c$.

This may not be unreasonable if we are dealing with the control of a bursty class-1 traffic.

Theorem 3.2 *Under the hypotheses of Theorem 3.1 and the additional assumption (A5), there exist jointly stationary processes A_1^* and A_2^* which are extremal in the sense that $\varphi(A_1^*, A_2^*) = \varphi^*$, with φ^* given by (3.11).*

Proof It is best to define A_1^* and A_2^* by a picture: see Figure 4. Also, for $i = 1, 2$ and $t \in \mathbb{R}$, define

$$X_i(t) = \sup_{-\infty < s \leq t} A_i(s_i, t] - \rho_i(t - s)$$

This is self-explanatory. Both processes are periodic with common period. The constant ξ_1 is defined by

$$\xi_1 := \frac{1}{\tau_0} \left(b - \frac{\pi_2 - c}{\pi_2 - \rho_2} \sigma_2 \right) = (\pi_2 - c) \frac{b}{\sigma_2} - \pi_2 + c.$$

Also,

$$\text{and } S_n - T_n = \max \left\{ \frac{b}{c}, \frac{\sigma_1}{\rho_1}, \frac{\sigma_2}{\rho_2} \right\}.$$

In this figure, we have taken $S_n - T_n = \sigma_1/\rho_1$. Again, to achieve stationarity we need to randomize the phase. (Notice that the same phase is used for both arrival processes.) \square

3.4 Dropping some assumptions: discussion. We have found a worst-case delay bound in the context of a queue with a constant service rate that is shared by multiple deterministically regulated sources. This bound was shown to be tight under the additional Assumption (A5) by the construction of an “extremal” pair of arrival processes.

Theorem 3.1 can be directly generalized beyond assumptions (A3) and (A4). For example, suppose we assume that $\pi_2 > c$ but we assume that $(\pi_2 - \rho_2)b \leq \sigma_2(\pi_2 - c)$. Also assume the background (A_2) arrivals are at rate π_2 over (S_0, R_0) and there are no class-1 arrivals over (S_0, R_0) . Thus, $R_0 - S_0 = b/(\pi_2 - c)$ and

$$X_2(R_0) = (\pi_2 - \rho_2) \frac{b}{\pi_2 - c} < \sigma_2.$$

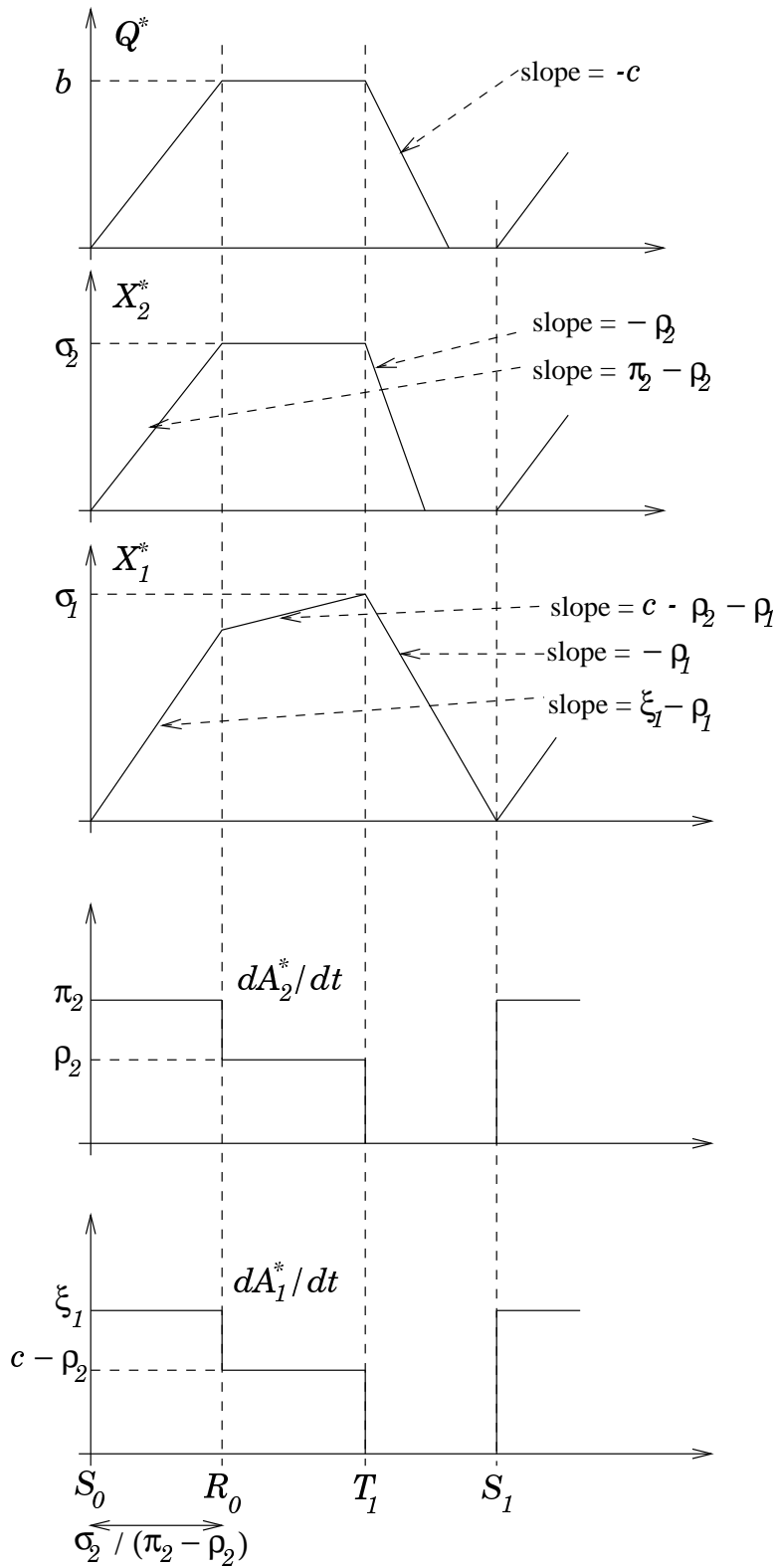


Figure 4 Periodic extremal arrival processes for a 2-class queue with dependent arrivals

So, Q can reach level b without the aid of class-1 arrivals and with residual slackness in the (σ_2, ρ_2) constraint. Consequently, we can arrange, in the worst-case, that *all* of the class-1 arrivals occur during the overload period (R_0, T_1) giving a trivial (but tight) upper bound of one on the quantity in (3.3).

4 Independent arrival processes

At the boundary of a communication network or at the input of an interior multiplexor, it may be natural to assume that arriving traffic sources are independent. In this section, we will find bounds on the overflow probability of a queue shared by multiple independent sources. Such a queue can be interpreted as the buffer part of an “access multiplexer” at a network boundary.

Consider a queue content process fed by n jointly stationary, mutually independent arrival processes A_1, \dots, A_n and constant service rate nc :

$$Q(t) = \sup_{-\infty < s \leq t} \left\{ \sum_{i=1}^n A_i(s, t] - nc(t-s) \right\}. \quad (4.1)$$

Also, for each $i = 1, 2, \dots, n$ define the following processes:

$$Q_i(t) = \sup_{-\infty < s \leq t} \{A_i(s, t] - c(t-s)\}. \quad (4.2)$$

The group of queues $\{Q_i\}_{i=1}^n$ was called the “virtual segregated system” in [25, 28]. Thus, Q_i is the queue content process of a system fed only by the i -th arrival process and served by the total service rate divided by n .

While we do not assume identical distribution for the A_i 's, we do, for the moment, assume identical shaping according to:

$$A_i(s, t] \leq \min\{\sigma + \rho(t-s), \pi(t-s)\} =: g(t-s), \quad s \leq t, \quad i = 1, \dots, n, \quad (4.3)$$

where $\sigma > 0$ and $\pi > c > \rho > 0$. As before, the abbreviations (2.4) shall be used.

4.1 Maximizing $\mathbf{E} \exp(\theta Q_i(0))$. We first find an upper-bound $\mathbf{E} \exp(\theta Q_i(0))$, for each i , over those processes A_i which are shaped as in (4.3). In fact, we shall be able to exactly maximize this functional as well as obtain the extremal process in this single-class case. The methods are analogous to the ones described in Section 2. Take $i = 1$.

Lemma 4.1 *Under the assumptions above, for all $\theta > 0$,*

$$\mathbf{E} e^{\theta Q_1(0)} \leq 1 - \alpha + \alpha \exp(\theta \alpha^{-1} (\sigma - (c - \rho)\gamma_\theta)) \quad (4.4)$$

where γ_θ is a constant depending only on the parameters $\sigma, \rho, \pi, c, \theta$ such that

$$\frac{\pi}{c} \cdot \frac{\sigma}{\pi - \rho} < \gamma_\theta < \frac{\sigma}{c - \rho}. \quad (4.5)$$

Furthermore, the bound (4.4) is tight because it is achieved by the periodic three-rate process depicted in Figure 5.

Proof Let time S_k , respectively V_k , be the beginning, respectively end, of the k -th busy period of Q_1 for $k \in \mathbb{Z}$. Let $\tau_k = V_k - S_k$ and $\tau'_k = S_{k+1} - V_k$, i.e., τ_k , respectively τ'_k , is the duration of the k -th busy, respectively idle, period of Q_1 ; see Figure 6. Letting \mathbf{P}^\dagger be the Palm probability of \mathbf{P} with respect to the points

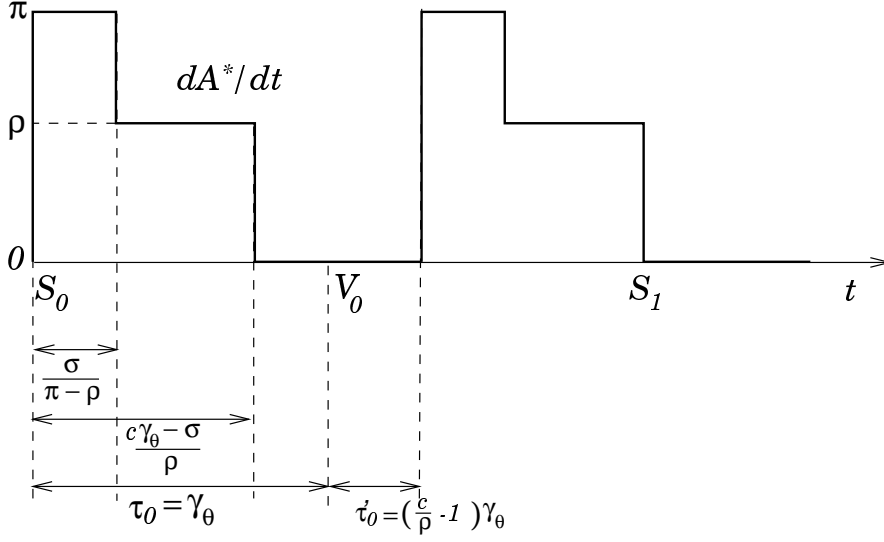


Figure 5 An arrival process that maximizes $Ee^{\theta Q_1(0)}$

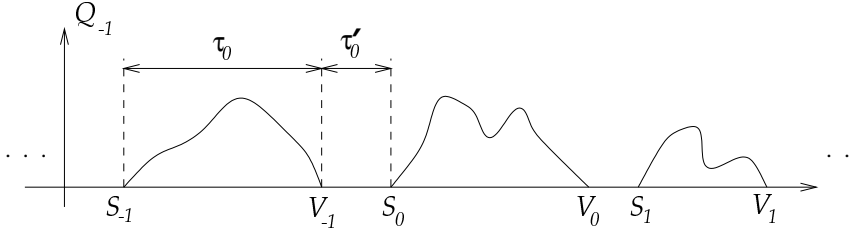


Figure 6 A sample path of the Q_1 process

$\{S_k, k \in \mathbb{Z}\}$, we have, by the Palm inversion formula:

$$\begin{aligned} Ee^{\theta Q_1(0)} &= \frac{E^\sharp \int_{S_0}^{V_0} e^{\theta Q_1(t)} dt + E^\sharp(S_1 - V_0)}{E^\sharp(S_1 - S_0)} \\ &= \frac{E^\sharp \int_{S_0}^{V_0} e^{\theta Q_1(t)} dt + E^\sharp \tau'_0}{E^\sharp \tau_0 + E^\sharp \tau'_0} = \frac{y + z}{1 + z} \end{aligned} \quad (4.6)$$

where

$$y = \frac{1}{E^\sharp \tau_0} E^\sharp \int_{S_0}^{V_0} e^{\theta Q_1(t)} dt, \quad (4.7)$$

$$z = \frac{E^\sharp \tau'_0}{E^\sharp \tau_0}, \quad (4.8)$$

and E^\sharp is expectation with respect to P^\sharp . Note that y is the P -expectation of $e^{\theta Q_1(0)}$ conditional on $Q_1(0) > 0$. Also note that the ratio that $(y + z)/(1 + z)$ is increasing in y and, since $y > 1$, the same ratio is decreasing in z . Thus, our objective is to maximize y and minimize z .

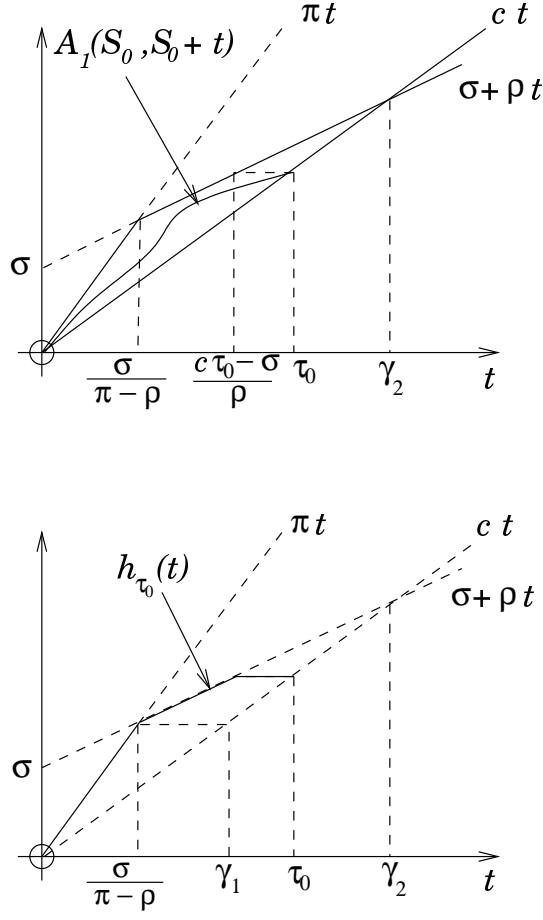


Figure 7 Cumulative arrival process and its bounding curve during a busy period

To minimize z , we first define

$$X_1(t) = \sup_{-\infty < s \leq t} \{A_1(s, t) - \rho(t - s)\}.$$

Note that $A_1(S_0, V_0) = c\tau_0$. Since the busy periods of Q_1 are entirely contained in those of X_1 ,

$$X_1(V_0) = X_1(S_0) + A_1(S_0, V_0) - \rho\tau_0 = X_1(S_0) + (c - \rho)\tau_0.$$

Also note that

$$X_1(S_1) \geq X_1(V_0) + A_1(V_0, S_1) - \rho\tau'_0 \geq X_1(V_0) - \rho\tau'_0 = X_1(S_0) + (c - \rho)\tau_0 - \rho\tau'_0.$$

Finally, since $\mathbb{E}^\dagger X_1(S_1) = \mathbb{E}^\dagger X_1(S_0)$, we have, by taking \mathbb{P}^\dagger -expectations in the above display, $0 \geq (c - \rho)\mathbb{E}^\dagger \tau_0 - \rho\mathbb{E}^\dagger \tau'_0$. Substituting in (4.8) we obtain

$$z \geq \frac{c}{\rho} - 1 = \frac{1}{\alpha} - 1. \quad (4.9)$$

To maximize y , first write

$$Q_1(t) = A_1(S_0, S_0 + t) - ct, \text{ for all } 0 \leq t \leq \tau_0 = V_0 - S_0, \quad (4.10)$$

because the process Q is positive on the interval (S_0, V_0) (see Figure 6). Then note that for all $t \in (0, \tau_0)$,

$$A_1(S_0, S_0 + t) \leq \min\{\sigma + \rho t, \pi t\} = g(t). \quad (4.11)$$

Observe that $A_1(S_0, V_0) = c\tau_0$, because the process Q is zero at the endpoints of (S_0, V_0) , and since $A_1(S_0, S_0 + t) \leq A_1(S_0, V_0)$, for all $t \in (0, \tau_0)$, we have a strengthening of (4.11):

$$A_1(S_0, S_0 + t) \leq \min\{\sigma + \rho t, \pi t, c\tau_0\}, \text{ for all } t \in (0, \tau_0). \quad (4.12)$$

Our goal is to obtain a tight upper bound for the exponent in the expression (4.7) for y . We find it convenient to introduce the the function

$$h_s(t) := g(t \wedge g^{-1}(cs)), \quad 0 \leq t \leq s. \quad (4.13)$$

It is easy to see that $h_{\tau_0}(t)$ is the right hand side of (4.12). So, if we further set

$$f(s) := \frac{1}{s} \int_0^s \exp(\theta(h_s(t) - ct)) dt, \quad (4.14)$$

we see that, for $\theta > 0$,

$$\begin{aligned} y &= \frac{1}{\mathbb{E}^\dagger \tau_0} \mathbb{E}^\dagger \int_0^{\tau_0} e^{\theta Q_1(t)} dt, \text{ because } \mathbb{P}^\dagger(S_0 = 0) = 1 \\ &= \frac{1}{\mathbb{E}^\dagger \tau_0} \mathbb{E}^\dagger \int_0^{\tau_0} e^{A_1(0,t) - ct} dt, \text{ due to (4.10)} \\ &\leq \frac{\mathbb{E}^\dagger \tau_0 f(\tau_0)}{\mathbb{E}^\dagger \tau_0}, \text{ due to definitions (4.14), (4.13), and inequality (4.12).} \end{aligned}$$

The trick now is to show that

$$f(\tau_0) \leq f(\gamma_\theta),$$

where γ_θ is a point at which f is maximal. This will be shown to be the case with γ_θ satisfying (4.5). We thus study, in the sequel, the properties of the function f . Let

$$\gamma_1 := \frac{\pi}{c} \cdot \frac{\sigma}{\pi - \rho}, \quad \gamma_2 := \frac{\sigma}{c - \rho}.$$

Observe that $\gamma_1 < \gamma_2$, because $c < \pi$. Using $g(t) = (\pi t) \wedge (\sigma + \rho t)$ in (4.13) we have the following expression for $h_s(t)$:

$$h_s(t) = (\pi t) \wedge (\sigma + \rho t) \wedge [(cs) \vee (\pi \rho^{-1}(cs - \sigma))] \wedge [(cs) \vee (\sigma + \rho \pi^{-1}cs)].$$

A rather tedious but straightforward procedure leads to the verification of:

$$h_s(t) = \begin{cases} \pi t, & \text{if } \{0 \leq s \leq \gamma_1, 0 \leq t \leq \frac{cs}{\pi}\} \text{ or } \{\gamma_1 \leq s \leq \gamma_2, 0 \leq t \leq \frac{\sigma}{\pi - \rho}\} \\ cs, & \text{if } \{0 \leq s \leq \gamma_1, \frac{cs}{\pi} \leq t \leq s\} \text{ or } \{\gamma_1 \leq s \leq \gamma_2, \frac{cs - \sigma}{\rho} \leq t \leq s\} \\ \sigma + \rho t, & \text{if } \gamma_1 \leq s \leq \gamma_2, \frac{\sigma}{\pi - \rho} \leq t \leq \frac{cs - \sigma}{\rho}. \end{cases}$$

Using this, we can compute $sf(s)$, as follows. First, for $0 \leq s \leq \gamma_1$,

$$sf(s) = \int_0^{cs/\pi} e^{\theta(\pi - c)t} dt + \int_{cs/\pi}^s e^{\theta c(s-t)} dt = \frac{\exp(\theta(\pi - c)\frac{c}{\pi}s) - 1}{\theta(\pi - c)\frac{c}{\pi}}.$$

Since $(e^u - 1)/u$ is increasing in $u \geq 0$, it follows that f is increasing on $[0, \gamma_1]$. Hence the maximum of f is located after γ_1 . Second, for $\gamma_1 \leq s \leq \gamma_2$,

$$\begin{aligned} sf(s) &= \int_0^{\frac{\sigma}{\pi-\rho}} e^{\theta(\pi-c)t} dt + \int_{\frac{\sigma}{\pi-\rho}}^{\frac{cs-\sigma}{\rho}} e^{\theta(\sigma-(c-\rho)t)} dt + \int_{\frac{cs-\sigma}{\rho}}^s e^{\theta c(s-t)} dt \\ &= \frac{e^{\theta\beta\sigma} - 1}{\theta(\pi-c)} + \frac{e^{\theta\beta\sigma}}{\theta(c-\rho)} - \frac{1}{\theta c} - \frac{1}{\theta\alpha^{-1}(c-\rho)} e^{\theta\alpha^{-1}(\sigma-(c-\rho)s)}. \end{aligned}$$

Therefore,

$$\frac{d}{ds}(sf(s)) = \exp(\theta\alpha^{-1}(\sigma - (c-\rho)s))$$

and, consequently,

$$\frac{d}{ds}f(s) = \frac{1}{s} \exp(\theta\alpha^{-1}(\sigma - (c-\rho)s)) - f(s). \quad (4.15)$$

At $s = \gamma_1$ we have

$$f'(\gamma_1) = \gamma_1^{-1}(e^{\theta\beta\sigma} - f(\gamma_1)).$$

But $f(\gamma_1) < e^{\theta\beta\sigma}$, because the mean of $\exp(\theta(h_{\gamma_1}(s) - cs))$ is less than its maximum. Hence $f'(\gamma_1) > 0$. At $s = \gamma_2$ we have

$$f'(\gamma_2) = \gamma_2^{-1}(1 - f(\gamma_2)).$$

But $f(\gamma_2) > 1$, because the mean of $\exp(\theta(h_{\gamma_2}(s) - cs))$ is more than its minimum. Hence $f'(\gamma_2) < 0$. These two observations allow us to conclude that f is maximized at some γ_θ between γ_1 and γ_2 , as asserted in (4.5). At this point, $f'(\gamma_\theta) = 0$, and from (4.15) we find

$$f(\gamma_\theta) = \exp(\theta\alpha^{-1}(\sigma - (c-\rho)\gamma_\theta)).$$

We conclude that

$$y \leq \frac{\mathbb{E}^i \tau_0 f(\tau_0)}{\mathbb{E}^i \tau_0} \leq f(\gamma_\theta). \quad (4.16)$$

Substituting the lower bound (4.9) for z and the upper bound (4.16) for y into (4.6), we arrive at (4.4) as desired. Finally, one can easily check that the arrival process of Figure 5 achieves this bound. \square

4.2 Bounding the shared queue. We can now prove a bound for the probability $\mathbb{P}(Q(0) \geq nb)$.

Theorem 4.2 *If $b \leq \beta\sigma$ then, for all n ,*

$$\mathbb{P}(Q(0) \geq nb) \leq \exp\left(-n \sup_{\theta > 0} F(\theta)\right)$$

where

$$F(\theta) := b\theta - \log(1 - \alpha + \alpha \exp(\theta\alpha^{-1}(\sigma - (c-\rho)\gamma_\theta))).$$

Proof From (4.1) and (4.2) we obtain

$$Q(0) \leq \sum_{i=1}^n \sup_{-\infty < t \leq 0} (A_i(t, 0) + ct) = \sum_{i=1}^n Q_i(0). \quad (4.17)$$

Using the Chernoff bound we get, for all $\theta > 0$,

$$\mathbb{P}(Q(0) \geq nb) \leq \frac{\mathbb{E} e^{\theta Q(0)}}{e^{\theta nb}}. \quad (4.18)$$

By (4.17), and the independence among the Q_i 's,

$$\mathbb{E}e^{\theta Q(0)} \leq \mathbb{E} \exp \left(\theta \sum_{i=1}^n Q_i(0) \right) = \prod_{i=1}^n \mathbb{E}e^{\theta Q_i(0)}$$

Since Lemma 4.1 can be applied to each of the $\mathbb{E}e^{\theta Q_i(0)}$, we get that

$$\mathbb{E}e^{\theta Q(0)} \leq (1 - \alpha + \alpha \exp(\theta \alpha^{-1}(\sigma - (c - \rho)\gamma_\theta)))^n.$$

Substituting this inequality into (4.18) and minimizing over $\theta > 0$ gives the desired result. \square

A looser bound can be obtained by using the fact that $Q_1(0) \leq \beta\sigma$ to get $y \leq e^{\theta\beta\sigma}$ in the proof of Lemma 4.1. This simpler bound can also be obtained by substituting γ_1 for γ_θ . With this simplification, $G(\theta) = b\theta - \log(1 - \alpha + \alpha e^{\beta\sigma\theta})$ is used instead of $F(\theta)$ in Theorem 4.2. Recall that $\beta = (\pi - c)/(\pi - \rho)$ and note that $G''(\theta) < 0$ for all $\theta > 0$, i.e., G is concave. Also, $G(0) = 0$ and $G'(0) = b - \alpha\beta\sigma$. Thus, $b - \alpha\beta\sigma > 0$ implies that $\sup_{\theta > 0} G(\theta) > 0$ (so that the bound using G is nontrivial).

4.3 The non-identically shaped case. It is easy to generalize Theorem 4 to the independent but not identically shaped sources case. Assume that A_i is shaped by its own

$$g_i(t) := \min\{\sigma_i + \rho_i t, \pi_i t\},$$

and let

$$\pi := \sum_{i=1}^n \pi_i, \quad \sigma := \sum_{i=1}^n \sigma_i, \quad \rho := \sum_{i=1}^n \rho_i.$$

Theorem 4.3 *If the A_i are stationary and independent, $0 < b < \frac{\pi - nc}{\pi - \rho}\sigma$ and $\pi > nc > \rho > 0$, then*

$$\mathbb{P}(Q(0) \geq b) \leq e^{-I_n}$$

with

$$I_n := \sup_{\theta > 0, \sum_i c_i = nc} \left\{ b\theta - \sum_{i=1}^n \log \left(1 - \alpha_i + \alpha_i \exp(\theta \alpha_i^{-1}(\sigma_i - (c_i - \rho_i)\gamma_{\theta,i})) \right) \right\},$$

where $\alpha_i = \rho_i/c_i$ and $\gamma_{\theta,i}$ is the quantity γ_θ defined in Lemma 4.1 using the parameters $\sigma_i, \rho_i, \pi_i, c_i$ and θ .

5 Comments and extensions

For large n , Theorem 4.2 can be sharpened by using the Bahadur-Rao theorem [11]. Also, a similar result for cell queuing delay of the aggregate flow (and, in the i.i.d. case, that of any particular source) can be obtained from the distributional Little's result, see, e.g., [23].

That on-off sources are not extremal in the independent multiple-source case was shown by simulation in [27]. The large deviations ($n \rightarrow \infty$) regime for the queue considered in Section 4 was discussed in [5, 9]. For the i.i.d. case, under some conditions on the arrival processes [5, 8],

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(Q(0) \geq nb) = - \inf_{t < 0} \sup_{\theta > 0} \{ b\theta - \log \mathbb{E} \exp(\theta(A_1(t, 0) + ct)) \}$$

This bound is *asymptotically* tighter than that of Theorem 4.2; indeed, for the i.i.d. case, the Chernoff bound (4.18) and (4.17) imply

$$\frac{1}{n} \log \mathbf{P}(Q(0) \geq nb) \leq - \sup_{\theta > 0} \inf_{t < 0} \{b\theta - \log \mathbf{E} \exp(\theta(A_1(t, 0) + ct))\}. \quad (5.1)$$

We note, however, that the bound on Theorem 4.2 holds for *all* n and that the bound given in Lemma 4.1 is tight. In Section 3.2 of [9], a symmetric 3-rate $(0, \rho, \pi)$ periodic process A_1 is shown to maximize (5.1).

References

- [1] Anantharam, V. and Konstantopoulos, T. (1994) Burst reduction properties of the leaky bucket flow control scheme in ATM networks. *IEEE Trans. Comm.* **42**, 3085-3089.
- [2] Anantharam, V. and Konstantopoulos, T. (1994) Optimality and interchangeability of leaky buckets in tandem. *Proc. 32nd Allerton Conference*, 235-244.
- [3] Anantharam, V. and Konstantopoulos, T. (1999) A methodology for the design of optimal traffic shapers in communication networks. *IEEE Trans. on Aut. Control*, to appear, March 1999.
- [4] Baccelli, F. and Brémaud, P. (1994) *Elements of Queueing Theory*. Springer-Verlag, New York.
- [5] Botvich, D.D., and Duffield, N.G. (1995) Large deviations, the shape of the loss curve, and economies of scale. *Queueing Systems* **20**, 293-320.
- [6] Bonatti, M. and Gaivoronski, A.A. (1994) Worst case analysis of ATM sources with application to access engineering of broadband multiservice networks. *Proc. International Teletraffic Congress (ITC-14)*, 559-569.
- [7] Chang, C.S. (1994) Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. Auto. Control*, **39**, 913-931.
- [8] C. Courcoubetis and R. Weber. (1996) Buffer overflow asymptotics for a switch handling many traffic sources. *J. Appl. Prob.*, **33**, 886-903.
- [9] Courcoubetis, C., Kelly, F., and Weber, R. (1998) Measurement-based usage charges in communications networks. *preprint* URL <http://www.statslab.cam.ac.uk/~frank/PAPERS/>
- [10] Cruz, R.L. (1991) A calculus for network delay, Part 1: Network elements in isolation. *IEEE Trans. Inform. Theory* **37**, 114-131.
- [11] Dembo, A. and Zeitouni, O. (1992) *Large Deviations Techniques and Applications*. Jones and Bartlett, Boston.
- [12] Doshi, B.T. (1994) Deterministic rule based traffic descriptors for broadband ISDN: worst case behavior and connection acceptance control. *Proc. International Teletraffic Congress (ITC-14)*, 591-600.
- [13] Duffield, N.G. and O'Connell, N. (1995) Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Cam. Phil. Soc.*, **118**, 363-374.
- [14] Elwalid, A. and Mitra, D. (1993) Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking* **1**, 329-343.
- [15] Elwalid, A., Mitra, D. and Wentworth, R.H. (1995) A new approach for allocating buffers and bandwidth to heterogeneous regulated traffic in an ATM node. *IEEE JSAC* **13**, 1115-1127.
- [16] Ganesh, A., Green, P., O'Connell, N. and Pitts, S. (1998) Bayesian network management. *Queueing Systems* **28**, 267-282.
- [17] Glynn, P.W. and Whitt, W. (1994) Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.* **31**.
- [18] Kelly, F.P. (1991) Effective bandwidths of multi-class queues. *Queueing Systems* **9**, 5-16.
- [19] Kesidis, G. and Konstantopoulos, T. (1999) Extremal shape-controlled traffic patterns in high-speed networks. *IEEE Trans. Comm.*, to appear. Also in: *Proc. IEEE CDC*, Tampa Bay, Florida.
- [20] Kesidis, G. (1999). *ATM Network Performance* (2nd Edition). Kluwer Academic Publishers, Boston.
- [21] Kesidis, G., Walrand, J. and Chang, C.S. (1993) Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Trans. Networking* **1**, 424-428.
- [22] Konstantopoulos, T. and Anantharam, V. (1995) Optimal flow control schemes that regulate the burstiness of traffic. *IEEE/ACM Trans. Networking* **3**, 423-432.

- [23] Konstantopoulos, T. and Last, G. (1998) On the dynamics and performance of stochastic fluid systems. URL:<http://fb1.math.nat.tu-bs.de:80/~schueler/mitarb/last.htm>.
- [24] Konstantopoulos, T., Zazanis, M. and de Veciana, G. (1997) Conservation laws and reflection mappings with an application to multiclass mean value analysis for stochastic fluid queues. *Stoch. Proc. Appl.* **65**, 139-146.
- [25] LoPresti, F., Zhang, Z.L., Towsley, D. and Kurose, J. (1997) Source time scale and optimal buffer/bandwidth trade-off for regulated traffic in an ATM node. *Proc. IEEE INFOCOM, Kobe, Japan*.
- [26] Mitra, D. and Morrison, J.A. (1995) Multiple Time Scale Regulation and Worst Case Processes for ATM Network Control. *Proc. IEEE CDC, New Orleans, LA*, p. 353-358.
- [27] Oechslin, P. (1997) Worst Case Arrivals of Leaky Bucket Constrained Sources: The Myth of the On-Off source. In *Proc. IFIP Fifth International Workshop on Quality of Service (IWQoS'97)*.
- [28] Rajagopal, S., Reisslein, M. and Ross, K.W. (1998) Packet multiplexers with adversarial regulated traffic. *Proc. IEEE INFOCOM'98, San Francisco*.
- [29] Roberts, J.W. (1994) Virtual spacing for flexible traffic control. *International Journal of Communication Systems* **7**, 307-318.
- [30] Saito, H. (1992) Call admission control in an ATM network using upper bound of cell loss probability. *IEEE Trans. Comm.* **40**, 1512-1521.
- [31] Turner, J. (1986) New directions in communication networks (or which way to the information age?). *IEEE Comm. Mag.* **24**, 8-15.