Journal of the Operations Research
Society of Japan
2004, Vol. 47, No. 4, 359-378

STATISTICAL MULTIPLEXING OF REGULATED SOURCES HAVING DETERMINISTIC SUBADDITIVE ENVELOPES

Kiwamu Nakamura Shigeo Shioda
Chiba University

(Received September 30, 2003; Revised May 14, 2004)

Abstract We study a single-server queue fed by various kinds of sources, each of which is constrained by a deterministic regulator, e.g., a token bucket. In particular, we derive bounds of virtual-waiting-time distribution only assuming that sources are stationary, statistically independent from each other, and have deterministic subadditive envelopes without using a specific traffic pattern. Based on the derived bounds, we investigate how large statistical multiplexing gain can be achieved when regulated sources share a common network resource. Numerical examples reveal that the regulated sources are more advantageous than Markov arrival processes or long-range-dependent sources with respect to the statistical multiplexing gain.

Keywords: Applied probability, queue, telecommunication

1. Introduction

In a typical computer network, each computer sends or receives data by using common network resources. Since each computer sends data intermittently, it is seldom that several computers simultaneously send a large amount of data and thus the shared network resources are fully occupied. This is the reason why the shared network resource can achieve much higher utilization than resources exclusively dedicated to a specific pair of computers. This resource-sharing effect is usually called *statistical multiplexing*.

While the Internet benefits largely from the statistical multiplexing, it usually does not give any quality-of-service (QoS) guarantees. The provision of QoS guarantees in the Internet has become a central topic of research for the last decade because congestion is widespread in today's Internet and QoS-aware applications like IP telephony are becoming major services. To this end, several QoS-guaranteeing mechanisms, including Intserv and Diffserv, have been discussed in the IETF.

Intserv [6] is a technology suitable for the deterministic QoS guarantee, which assures a given worst-case delay bound or no loss in the Internet. Although the deterministic QoS guarantee is the most stringent and its realization might be possible by deterministically assign the network resource to each data flow [11, 12, 23], it cannot benefit from the statistical multiplexing at all and thus the utilization of network resource should be too low. Diffserv [3, 18] is another solution for introducing QoS to the Internet: it is suitable for realizing the statistical QoS guarantee of the form such as $P[\text{delay} > d_{target}] < \epsilon$, where d_{target} is a target-delay bound and ϵ typically ranges from 10^{-3} to 10^{-9}. The statistical QoS guarantee can usually achieve much higher utilization than the deterministic QoS guarantee because the former can take advantage of the statistical multiplexing. We should, however, note that there is a general belief that any kinds of QoS guarantees will necessitate operating the Internet at very low utilization because of the bursty nature of the Internet traffic.

Since it is not often feasible to obtain reliable statistical characteristics of sources, recent research on the statistical QoS has attempted to assess the QoS of the form like $P[\text{delay} > d_{target}]$ without making assumptions on statistical properties of sources. In particular, a number of studies have been made to devise the techniques for assessing the QoS by only assuming that the amount of traffic from each source is constrained by a deterministic regulator [4, 8, 16, 19–21, 24, 25, 29]. For example, Elwalid *et al.* [16] and LoPresti *et al.* [24] studied packet-loss probability in a multiplexer where a large number of regulated sources are multiplexed. They assumed that traffic from each regulated source follows a periodic on-off pattern. Several works (for example [25]), however, revealed that the periodic on-off pattern does not maximize QoS violations: another class of patterns may be even worse. Thus, the QoS evaluation based on the periodic on-off pattern has a weakness in the sense that it does not generally give conservative evaluation results. Kesidis *et al.* [19, 20] and Shioda [25] addressed the problem of finding the worst traffic pattern among all possible patterns that are constrained by a given regulator. Boorstyn *et al.* [4] derived the *effective envelopes* of superposition of regulated sources, each of which has a deterministic subadditive envelope, without using a specific traffic pattern. Kesidis and Konstantopoulos [21] derived the bound of the workload distribution for superposition of independent homogeneous regulated sources. Chang, Chiu, and Song [8] considered the same problem as Kesidis and Konstantopoulos in a discrete-time model and derived a different bound of the workload distribution. Vojnović and Le Boudec [29] extended their results to the cases where multiplexed regulated sources have different envelopes from each other (that is, heterogeneous cases) in a continuous time setting. They also considered a case where the outgoing links from network nodes do not have constant but time-varying capacities. (If network nodes use some scheduling algorithms like weighted-fair queueing or deficit round robin, then the network node do not offer the constant service rate at each instant of time [27].)

The aim of this paper is to investigate how large statistical multiplexing gain can be achieved when regulated sources share a common network resource. For this purpose, we derive two bounds (one is for a discrete-time setting and the other is for a continuous time setting) of the delay distribution $P[\text{delay} > d]$ of a single-server queue fed by heterogeneous regulated sources. We derive the bounds only assuming that sources are stationary, statistically independent from each other, and have deterministic subadditive envelopes without using a specific traffic pattern. The derived bounds have explicit expressions so that they can be easily calculated.

Our results include a continuous-time and heterogeneous extension of the bound in Chang, Chiu, and Song [8]*. Note that the continuous-time model is very important for analyzing variable-length-packet networks like the Internet. The extension of the discrete-time result to the continuous-time one is, however, not trivial because simply letting the length of time unit in a discrete time model be zero does not yield the appropriate bound in the corresponding continuous-time model. In addition to this, our bound is applicable to cases where network nodes do not offer the constant service rate. This problem setting is same as that in Vojnović and Le Boudec [29], but our bound is tighter than their bound. We also numerically show that our bound is significantly tighter than that in Kesidis and Konstantopoulos [21].

The remainder of this paper is organized as follows. In Section 2, we describe the system model and assumptions on the traffic characteristics used in the analysis. In Section 3,

*Although Chang *et al.* [8] also showed the bound for heterogeneous regulated sources, they did not show how it is derived.

we derive bounds of the virtual-waiting-time distribution of a single-server queue in both discrete- and continuous-time models. In Section 4, we use the derived bounds to investigate the statistical multiplexing gain of regulated sources. In Section 5, we numerically compare the proposed bound with other existing bounds or the exact bound to investigate the tightness of our bound. Using the proposed bounds, we also show how large statistical-multiplexing gain can be obtained for regulated traffic sources. In section 6, we present concluding remarks of our work.

2. Preliminaries

We first describe the system model used in this paper. Consider packet arrivals to an output buffer in a network node. As shown in Figure 1, the arrivals from sources are policed by a regulator, switched to an outgoing link, and inserted into the output buffer dedicated to the outgoing link. The data in the output buffer is transferred to the outgoing link with time-varying (stochastic) rate. We focus on the delay due to waiting in the output buffer, which can be regarded as a single server queue.

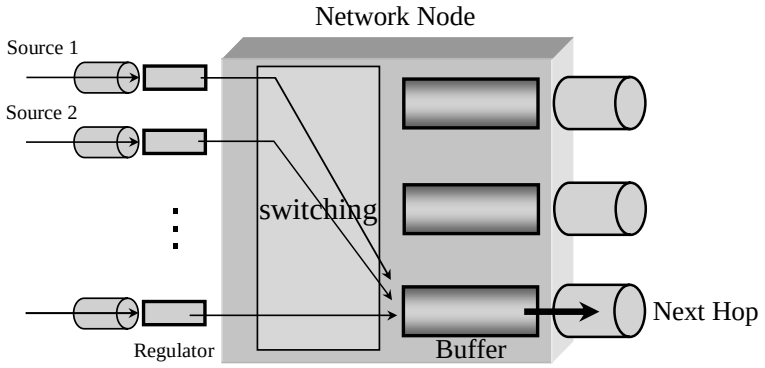

Figure 1: Packet switching in a network node

We assume that the sources are classified into K classes according to their regulator characterization, where all regulated sources in class k $(1 \le k \le K)$ have the same regulators. There are N_k class-k sources. Let $A_j^{(k)}(t, t+\tau)$ be the amount of data (say in bit) arriving during $(t, t+\tau]$ from source j in class k. We assume that $A_j^{(k)}(t, t+\tau)$ has the following characteristics:

(1) Subadditive bound: $A_j^{(k)}(t, t+\tau)$ is regulated by a deterministic subadditive† envelope [4, 9] such that

$$A_j^{(k)}(t, t+\tau) \le \alpha^{(k)}(\tau) < \infty$$

for all t and $\tau > 0$.

(2) Stationarity: $A_j^{(k)}(t, t+\tau)$ is stationary. Then, for example,

$$P[A_j^{(k)}(t, t+\tau) \le x] = P[A_j^{(k)}(t', t'+\tau) \le x]$$

for all $t, t' > 0$.

†A function $f(x)$ is called subadditive (superadditive) if

$$f(x+y) \le (\ge) f(x) + f(y).$$

(3) Independence: For all $i \neq j$, $A_i^{(k)}$ and $A_j^{(k)}$ are stochastically independent. Similarly, if $k \neq l$, $A_i^{(k)}$ and $A_j^{(l)}$ are stochastically independent for all i, j.

The traffic from a source regulated by a dual token bucket [6], which is the most popular regulator for peak- and average-rate enforcements, has the following deterministic subadditive envelope:

$$\alpha^{(k)}(\tau) = \min\{P^{(k)}\tau, \sigma^{(k)} + \rho^{(k)}\tau\},$$

where $P^{(k)}$ is the peak traffic rate, $\rho^{(k)}$ is the average traffic rate, and $\sigma^{(k)}$ is a burst size parameter.

We also let $B(t, t+\tau)$ denote the maximum amount of data that can be transferred from the buffer during $(t, t+\tau]$. Note that $B(t, t+\tau)$ is a random variable. We assume that $B(t, t+\tau)$ has the following characteristics:

(4) Superadditive lower bound: $B(t, t+\tau)$ has a deterministic superadditive lower bound such that

$$B(t, t+\tau) \geq \beta(\tau)$$

for all $t, \tau > 0$. The lower bound $\beta(\tau)$ is usually called *service curve* [1, 23, 29].

(5) Stability condition:

$$\lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{K} N_k \alpha^{(k)}(t) < \lim_{t\to\infty} \frac{\beta(t)}{t}. \tag{2.1}$$

(Subadditivity of $\alpha^{(k)}(t)$ and superadditivity of $\beta(t)$ guarantee the existence of limits in (2.1) [4].)

If the buffer has a constant service rate C, then the service curve $\beta(\tau)$ is given by $C\tau$. If the data in the buffer is served by a latency-rate server with service rate C and latency e, then the service curve $\beta(\tau)$ is given by $C\max\{\tau - e, 0\}$ [27, 29]. (Several well-known scheduling algorithms, such as Weighted Fair Queueing, VirtualClock, and Deficit Round Robin, belong to the class of latency-rate servers.)

The following lemma provides a bound on the moment generating function of $A_j^{(k)}(0,t)$, which will be used to prove the results in the next section. Concerning the proof, please see [4, 25].

Lemma 2.1. *If $A_j^{(k)}(t, t+\tau)$ satisfies assumptions (1)-(3) explained above, then*

$$E[e^{\theta A_j^{(k)}(0,t)}] \leq 1 + \frac{\rho^{(k)} t}{\alpha^{(k)}(t)}(e^{\theta\alpha^{(k)}(t)} - 1),$$

where

$$\rho^{(k)} \stackrel{\text{def}}{=} \lim_{t\to\infty} \frac{\alpha^{(k)}(t)}{t}.$$

Remark 2.1. Condition (5) provides a kind of queue-stability condition. To see this, first observe that from conditions (1), (2) and (6)

$$\begin{aligned} \sum_{k=1}^{K} N_k E[A^{(k)}(0,1)] &\leq \sum_{k=1}^{K} N_k \rho^{(k)} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{K} N_k \alpha^{(k)} \\ &< \lim_{t\to\infty} \frac{\beta(t)}{t}. \end{aligned} \tag{2.2}$$

If the data in the buffer is served with deterministic service rate C (that is, $\beta(t) = Ct$), then (2.2) becomes

$$\sum_{k=1}^{K} N_k E[A^{(k)}(0,1)] < C,$$

which is a well-known stability condition of queues.

3. Upper Bound of Virtual-Waiting-Time Distribution

3.1. A discrete time model

We would like to start by considering a discrete-time model where data periodically arrives and is transferred at time nT where n is an arbitrary integer (Figure 2). (We assume that, at every data-arrival-transfer epoch, data first arrives at the buffer and then data in the buffer is transferred.) The amount of data that arrives at time nT from source j in class k is equal to $A_j^{(k)}((n-1)T, nT)$. Let $\tilde{D}^{(T)}(\vec{N})$ denote the steady-state virtual-waiting time in the discrete-time model when the number of sources of each class is $\vec{N} \stackrel{\text{def}}{=} (N_1, N_2, \cdots, N_K)$ and the length of time unit is T. The following result is an extension of the bound in Chang, Chiu, and Song [8] to the cases where network nodes do not offer the constant service rate but the service curve $\beta(\tau)$.

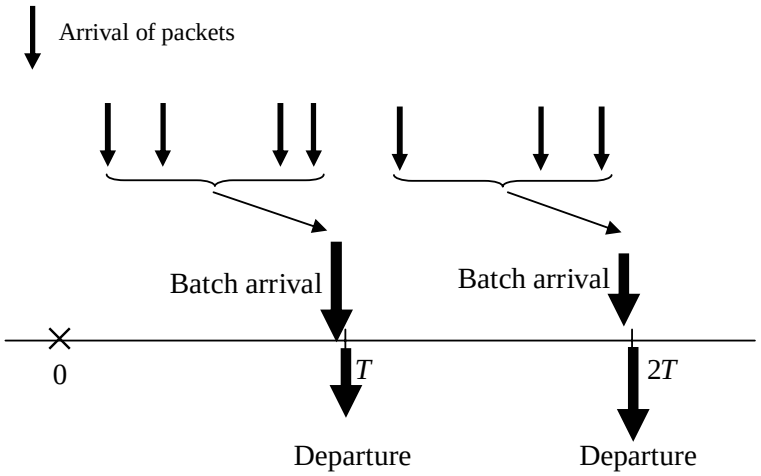

Figure 2: Discrete time model

Theorem 3.1. *If* $\sum_{k=1}^{K} N_k \rho^{(k)} t < \beta(t) + \beta(d)$ *for all* t, *then*

$$P[\tilde{D}^{(T)}(\vec{N}) > d] \leq \sum_{n=1}^{n_{max}(T,d,\vec{N})} \prod_{k=1}^{K} \left\{ \left(\frac{x_k(n,T)}{y_k(n,T,d,\vec{N})} \right)^{y_k(n,T,d,\vec{N})} \times \left(\frac{1-x_k(n,T)}{1-y_k(n,T,d,\vec{N})} \right)^{1-y_k(n,T,d,\vec{N})} \right\}^{N_k}, \qquad (3.3)$$

where

$$n_{max}(T,d,\vec{N}) \stackrel{\text{def}}{=} \sup\{n : \sum_{k=1}^{K} N_k \alpha^{(k)}(nT) > \beta(nT) + \beta(d)\},$$

$$x_k(n,T) \stackrel{\text{def}}{=} \frac{n\rho^{(k)}T}{\alpha^{(k)}(nT)}, \quad y_k(n,T,d,\vec{N}) \stackrel{\text{def}}{=} \frac{\gamma_k(n,T,\theta^*(n,T,d,\vec{N}))}{\alpha^{(k)}(nT)},$$
$$\gamma_k(n,T,\theta) \stackrel{\text{def}}{=} \frac{\alpha^{(k)}(nT)}{1-(1-\frac{\alpha^{(k)}(nT)}{n\rho^{(k)}T})e^{-\theta\alpha^{(k)}(nT)}},$$

and $\theta^*(n,T,d,\vec{N})$ *is the unique solution in* $(0,\infty)$ *to the following equation of* θ*:*

$$\beta(nT)+\beta(d) = \sum_{k=1}^{K} N_k\gamma_k(n,T,\theta). \tag{3.4}$$

Proof. See Appendix A.1. □

Remark 3.1. When $K=1$, the bound of virtual-waiting-time distribution (3.3) is expressed in the following simple form:

$$P[\tilde{D}^{(T)}(N)>d] \le \sum_{n=1}^{n_{max}(T,d,N)} \left\{ \left(\frac{x(n,T)}{y(n,T,d,N)}\right)^{y(n,T,d,N)} \left(\frac{1-x(n,T)}{1-y(n,T,d,N)}\right)^{1-y(n,T,d,N)} \right\}^N.$$

$$n_{max}(T,d,N) = \sup\{n : N\alpha(nT) > \beta(nT)+\beta(d)\},$$
$$x(n,T) = \frac{n\rho T}{\alpha(nT)}, \quad y(n,T,d,N) = \frac{\beta(nT)+\beta(d)}{N\alpha(nT)}.$$

Remark 3.2. Thanks to the stability condition (6), $n_{max}(T,d,\vec{N})$ is finite.
Remark 3.3. The solution of (3.4) can be easily found because the right hand side of (3.4) is increasing in θ.

3.2. A continuous-time model

Using the result in a discrete-time model, we derive the bound of the steady-state virtual-waiting-time distribution in a continuous-time model. Let $D(\vec{N})$ denote the steady-state virtual-waiting time in a continuous-time model when the number of sources of each class is $\vec{N}$.

Theorem 3.2. *If* $\sum_{k=1}^{K} N_k\rho^{(k)}t < \beta(t)+\beta(d)$ *for all* t*, then*

$$P[D(\vec{N})>d]$$
$$\le \sum_{n=1}^{\hat{n}_{max}(T,d,\vec{N})} \prod_{k=1}^{K} \left\{ \left(\frac{x_k(n,T)}{\hat{y}_k(n,T,d,\vec{N})}\right)^{\hat{y}_k(n,T,d,\vec{N})} \left(\frac{1-x_k(n,T)}{1-\hat{y}_k(n,T,d,\vec{N})}\right)^{1-\hat{y}_k(n,T,d,\vec{N})} \right\}^{N_k} \tag{3.5}$$

for all $T>0$*, where*

$$\hat{n}_{max}(T,d,\vec{N}) \stackrel{\text{def}}{=} \sup\{n : \sum_{k=1}^{K} N_k\alpha^{(k)}(nT) > \beta((n-1)T)+\beta(d)\},$$
$$\hat{y}_k(n,T,d,\vec{N}) \stackrel{\text{def}}{=} \frac{\gamma_k(n,T,\hat{\theta}^*(n,T,d,\vec{N}))}{\alpha^{(k)}(nT)},$$

and $\hat{\theta}^*(n,T,d,\vec{N})$ *is the unique solution in* $(0,\infty)$ *to the following equation of* θ*:*

$$\beta((n-1)T)+\beta(d) = \sum_{k=1}^{K} N_k\gamma_k(n,T,\theta).$$

Proof. See Appendix A.2. □

Remark 3.4. Simply letting $T \to 0$ in Theorem 3.1 does not yield the result in Theorem 3.2 because the summation in the right-hand side of (3.3) becomes infinite when letting $T \to 0$.

Remark 3.5. For homogeneous cases, letting $t_k = kT$ and $K = t/T$ in Theorem 3 in Vojnović and Le Boudec [29], which is the continuous-time version of the bound in Chang, Chiu, and Song [8], yields the same bound as (3.5). For heterogeneous cases, however, (3.5) yields tighter bound than Theorem 4 (heterogeneous version of Theorem 3) in [29] because the latter applies the Hoeffding's inequalities to derive the bound. (Also see, section 5.1.)

Remark 3.6. The choice of T affects the tightness of the bound. We numerically found that letting $T = d/2$ in (3.5) yields reasonably good bounds for almost all cases.

4. Statistical Multiplexing Gain Due to the Large Numbers of Superposition

In this section, we analyze the statistical multiplexing due to the large numbers of superposition when the transmission rate is scaled with the number of multiplexed sources. To simplify the analysis, we assume that $N_k = r_k N$ where $\sum_{k=1}^K r_k = 1$ and that $\beta(t) = \sum_{k=1}^K c_k N_k t = \sum_{k=1}^K c_k r_k N t$. Under these assumptions, we focus on the behavior of $D(\vec{N})$ when N increases while $\vec{c} \stackrel{\text{def}}{=} (c_1, \cdots, c_K)$ and $\vec{r} \stackrel{\text{def}}{=} (r_1, \cdots, r_K)$ are fixed. We obtain the following main result:

Theorem 4.1. *If* $\vec{c} > \vec{\rho}$ $(\stackrel{\text{def}}{=} (\rho_1, \cdots, \rho^{(K)}))^{\ddagger}$

$$P[D(\vec{N}) > d] \le \hat{n}_{max}(T, d, \vec{r}) e^{-\eta(T,d,\vec{r})N}, \tag{4.6}$$

where

$$\eta(T, d, \vec{r}) \stackrel{\text{def}}{=} -\sum_{k=1}^K r_k \log \left\{ \left(\frac{x_k(n^*, T)}{\hat{y}_k(n^*, T, d, \vec{r})} \right)^{\hat{y}_k(n^*,T,d,\vec{r})} \left(\frac{1 - x_k(n^*, T)}{1 - \hat{y}_k(n^*, T, d, \vec{r})} \right)^{1-\hat{y}_k(n^*,T,d,\vec{r})} \right\},$$

$$\hat{y}_k(n, T, d, \vec{r}) \stackrel{\text{def}}{=} \frac{\gamma_k(n, T, \hat{\theta}^*(n, T, d, \vec{r}))}{\alpha^{(k)}(nd)},$$

$$\hat{n}_{max}(T, d, \vec{r}) \stackrel{\text{def}}{=} \sup\{n : \sum_{k=1}^K r_k \alpha^{(k)}(nT) > ((n-1)T + d)c\}$$

Here, n^* *is the integer that maximizes*

$$\prod_{k=1}^K \left\{ \left(\frac{x_k(n, T)}{\hat{y}_k(n, T, d, \vec{r})} \right)^{\hat{y}_k(n,T,d,\vec{r})} \left(\frac{1 - x_k(n, T)}{1 - \hat{y}_k(n, T, d, \vec{r})} \right)^{1-\hat{y}_k(n,T,d,\vec{r})} \right\}^{N_k}$$

in $[0, \hat{n}_{max}(T, d, \vec{r})]$ *and* $\hat{\theta}^*(n, T, d, \vec{r})$ *is the unique solution to the following equation:*

$$((n-1)T + d)c = \sum_{k=1}^K r_k \gamma_k(n, T, \theta), \quad c \stackrel{\text{def}}{=} \sum_{k=1}^K c_k r_k.$$

Proof. See Appendix A.3. □

‡Here, we let $>$ denote coordinatewise ordering for real vector: for $\vec{x} = (x_1, \cdots, x_m)$ and $\vec{y} = (y_1, \cdots, y_m)$, $\vec{x} > \vec{y}$ if $x_i > y_i$ for $i = 1, \cdots, m$.

Remark 4.1. A exponential delay bound similar with (4.6) also holds for Markov arrival processes [7, 13, 15]. In this sense, the regulated sources achieve, at least, the same level of statistical multiplexing gain as Markov arrival processes.

Remark 4.2. Relationship (4.6) is closely related to the *large-superposition asymptotics* or *economies of scale* formula [5, 10, ?, 15]

$$\lim_{N\to\infty} N^{-1}\log P[D(cN,\vec{N})>d] = -\eta(d)N, \quad \text{or} \quad P[D(cN,\vec{N})>d] \approx e^{-\eta(d)N}, \tag{4.7}$$

where $\eta(d)$ is usually called the *shape function*. Botvich *et al.* [5] and Duffield [14] have shown that a wide variety of sources including some long-range-dependent sources satisfy this asymptote. Note that relationship (4.6) is stronger than the large-superposition asymptote (4.7): that is, (4.7) always holds for sources that satisfy (4.6).

Remark 4.3. In general, the shape function has the following asymptote:

$$\eta(d) \approx \delta d^{\epsilon_1} + \nu d^{\epsilon_2},$$

where $\epsilon_1 = 1$ and $\epsilon_2 = 0$ for Markov arrival processes and $0 < \epsilon_1, \epsilon_2 < 1$ for general long-range-dependent sources [14]. In other words, $\eta(d)$ is linear for Markovian sources while $\eta(d)$ is concave for general long-range-dependent sources. As will be shown in the next section, our numerical examples have revealed that the shape function of regulated sources is a convex function of d. The behavior of the shape functions of these sources is schematically drawn in Figure 3.

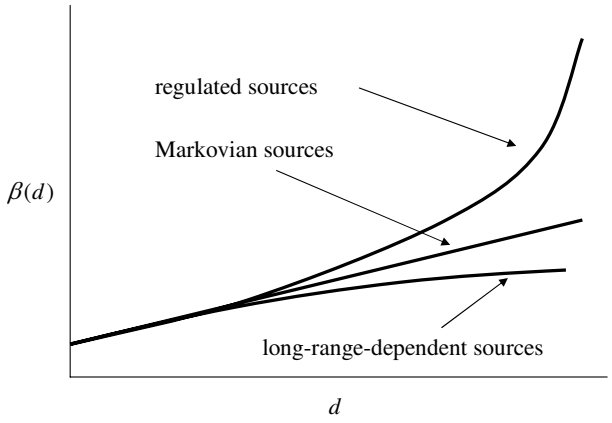

Figure 3: Behavior of the shape functions

This finding indicates that, as d increases, $P[delay > d]$ when regulated sources are multiplexed converges to 0 much faster than that when Markovian or long-range-dependent sources are multiplexed. In other words, if delay target is so large, the regulated sources could have larger multiplexing gain compared with Markovian and long-range-dependent sources.

5. Numerical Examples

5.1. Comparison with existing bounds

We numerically compared the proposed bound (3.5) with other existing bounds when a large number of IP-telephony sources are multiplexed. We consider two low-bit-rate voice-coding algorithms: G.723.1 and G.729 [22].

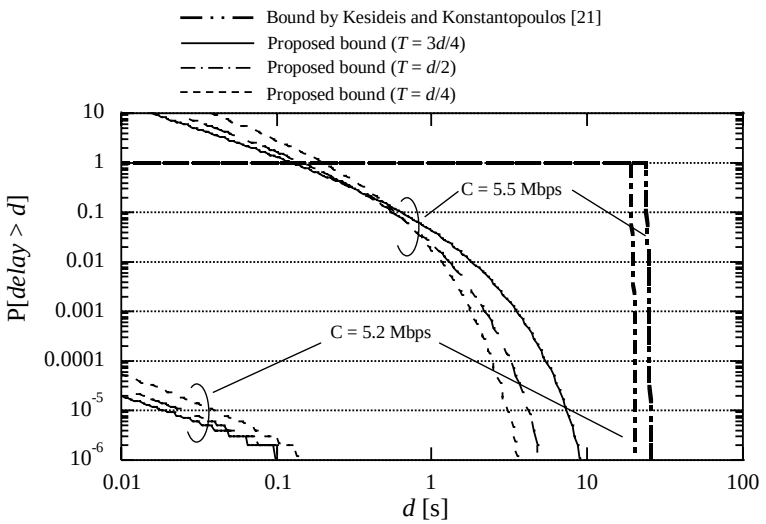

Figure 4: Comparison of bounds: homogeneous case

In numerical examples, we assumed that the bit rate of G.723.1 is 6.4 kbps. Since frames of G.723.1 are generated every 30 ms, the frame size of G.723.1 is $6.4 \times 30/8 = 24$ byte during a talkspurt, while the frame size during a silence period is 4 byte. Each IP packet is constructed from a single frame of G.723.1 with the IP/UDP/RTP header. We assumed that the size of the IP/UDP/RTP header is compressed into 4 byte by the header compression technique. Thus, the total size of an IP packet is 28 byte (in a talkspurt) or 8 byte (in a silence period). Now, let δ denote talkspurt activity. The average bit rate of a single IP-telephony source is then given by

$$\rho(\delta) = \delta \times 7.5 + (1 - \delta) \times 2.1 \text{ kbps} \tag{5.8}$$

because the bit rate in a talkspurt is $28 \times 8/30 = 7.5$ kbps and the one in a silence period is $8 \times 8/30 = 2.1$ kbps.

We first evaluated the supplementary delay distribution $P[delay > d]$ when a thousand of IP-telephony sources of G.723.1 are multiplexed. For this purpose, we assume that the traffic from a IP-telephony source of G.723.1 is transparent to a dual token bucket whose peak bit rate is 7.5 kbps, average bit rate is $\rho(\delta)$, and bucket size is

$$\sigma(\delta) = \{7.5 - \rho(\delta)\} \times 60 \quad \text{kbit.}$$

The bucket whose size is $\sigma(\delta)$ can store the data when the talkspurt lasts 60 seconds. Since the talkspurt duration usually ranges from 100 ms to 500 ms [26, 30], most of IP telephony sources should be transparent to the dual token bucket explained above. In Figure 4, we show the supplementary delay distribution $P[delay > d]$ for various values of T when $\delta = 0.5$ and the data in the buffer is transferred with constant service rate of 5.2 Mbps or 5.5 Mbps. We find that the proposed bound is significantly tighter than the bound in Kesidis and Konstantopoulos [21]. (Note that the bound in Kesidis and Konstantopoulos [21] is essentially the same as Theorem 1 in Vojnović and Le Boudec [29].) Similar findings were reported in [8, 29], and we now confirm this for IP-telephony sources. We also found that there is no optimal value for T so that $P[delay > d]$ becomes smallest for all d.

Next, we evaluated the supplementary delay distribution $P[delay > d]$ when IP-telephony sources coded by G.729 and those coded by G.723.1 are multiplexed together. The bit rate

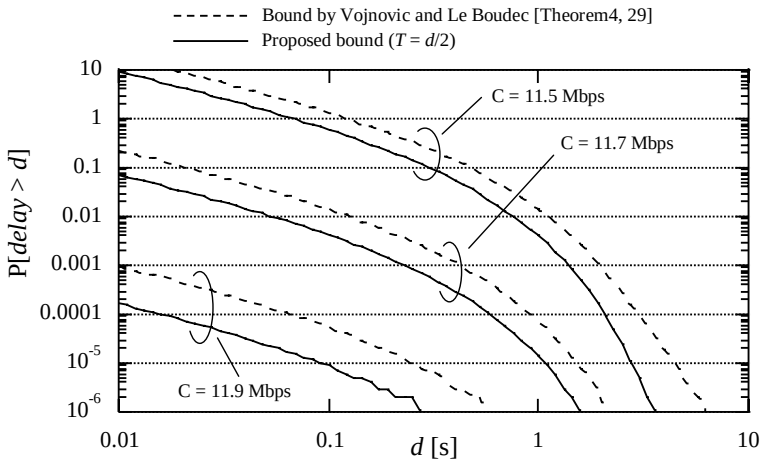

Figure 5: Comparison of bounds: heterogeneous case

of G.729 coding is 8 kbps. Note that, in G.729 coding, a ten-byte frame is generated every 10 ms during a talkspurt while a two-byte frame is generated every 20 ms during a silence period. Since an IP packet is constructed from two frames of G.729 with the IP/UDP/RTP header during a talkspurt, the total size of IP packet is 24 byte during a talkspurt. The IP packet during a silence period is made of a two-byte frame of G.729 and the IP/UDP/RTP header, so its size is 6 byte. The average bit rate of a single IP-telephony source is given by

$$\rho(\delta) = \delta \times 9.6 + (1 - \delta) \times 2.4 \text{ kbps} \tag{5.9}$$

because bit rate in a talkspurt is $24 \times 8/20 = 9.6$ kbps and that in a silence period is $6 \times 8/20 = 2.4$ kbps. Then, we also assume that the traffic from a IP-telephony source of G.729 is constrained by a dual token bucket whose peak bit rate is 9.6 kbps, average bit rate is $\rho(\delta)$, and bucket size is

$$\sigma(\delta) = \{9.6 - \rho(\delta)\} \times 60 \quad \text{kbit}.$$

Figure 5 shows the results when $\delta = 0.5$, $T = d/2$, and the data in the buffer is transferred with constant service rate of 11.5 Mbps, 11.7 Mbps or 11.9 Mbps. The number of G.729 sources and that of G.723.1 sources are both a thousand. We find that the proposed bound is tighter than Theorem 4 in Vojnović and Le Boudec [29].

5.2. Tightness of proposed bound

We evaluate the tightness of the proposed bound (3.5) when a thousand of IP-telephony sources coded by G.723.1 are multiplexed. We assume that the data is not compressed even during silence periods so that 28-byte-length packets periodically arrive every 30 ms from each source. Such a periodic source has the subadditive envelope such that $\alpha(\tau) = P\tau + L$, where P is the peak rate (the inverse of the packet-interarrival time) and L is the packet length. We also assume that IP telephony sources share the outgoing link with nonreal-time traffic. The packets of IP telephony sources are assumed to have nonpreemptive priority over the nonreal-time traffic. Note that, for such a case, the service curve for IP telephony sources is given by $\beta(\tau) = \min\{C\tau - L_{nonreal}, 0\}$ where C is the outgoing-link capacity and $L_{nonreal}$ is the length of nonreall-time packet. Iida *et al.* [17] derived the exact bound of delay distribution when periodic sources are served with nonpreemptive priority discipline, so we use their result for evaluating tightness.

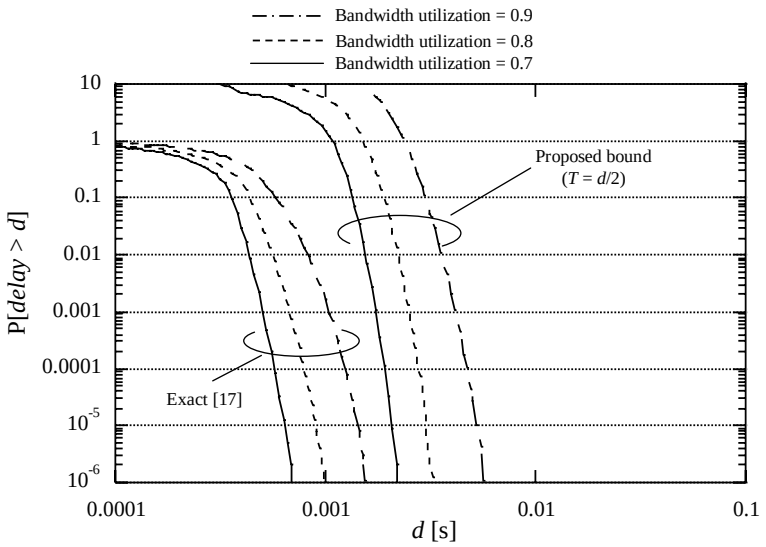

Figure 6: Comparison of the proposed bound with the exact bound [17]: periodic sources

In Figure 6, we compare the proposed bound of supplementary delay distribution with the exact bound in Iida *et al.* [17] for various bandwidth utilization. (The bandwidth utilization is defined as a ratio of the sum of peak rates of IP telephony sources to the outgoing-link capacity. In numerical experiments, we let $L_{nonreal} = 500$ byte.) In Table 1, we summarize the 99.9-percentile delay obtained by the proposed and the exact bound. Although there is some gap between the proposed and exact bound, we can conclude that the proposed delay bound is reasonably tight enough for practical use.

Table 1: 99.9 percentile delay

Bandwidth utiliztaion	Proposed formula [ms]	Exact formula [17] [ms]
0.7	1.7	0.50
0.8	2.5	0.67
0.9	4.1	1.0

5.3. Statistical multiplexing gain

Finally, we numerically investigate the statistical multiplexing gain when a large number of IP-telephony sources are multiplexed. First, we consider the case where a number of IP telephony sources coded by G.723.1 are multiplexed. We should note that the bandwidth utilization of the link dedicated to a single source is $\rho(\delta)/7.5$: for example, the bandwidth utilization is 0.64 when $\delta = 0.5$. If the bandwidth resource is shared by several sources, the bandwidth utilization should be larger than $\rho(\delta)/7.5$ thanks to the statistical multiplexing. Furthermore, the bandwidth utilization should become larger as the number of multiplexed sources increases. With this in mind, based on the delay bound (3.5), we evaluated the bandwidth utilization by changing the number of multiplexed sources when the target statistical QoS is given by $P[delay > d_{target}] < \epsilon$.

In Figure 7, we show the relationship between the number of multiplexed sources and the bandwidth utilization when $d_{target} = 20$ ms and $\delta = 0.5$. The 90%-bandwidth utilization

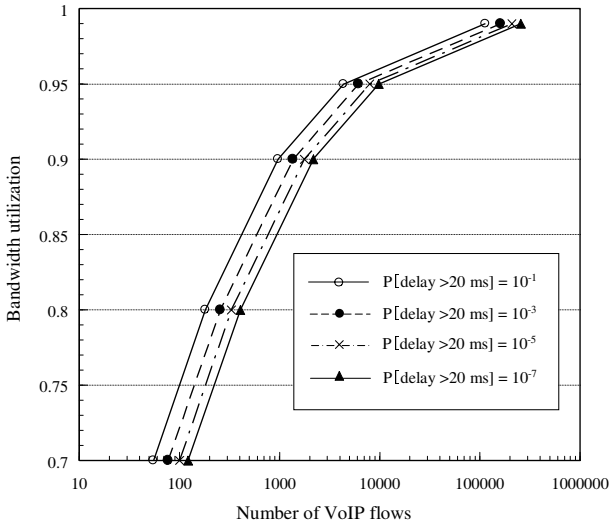

Figure 7: Relationship between the number of multiplexed sources and bandwidth utilization: G.723.1

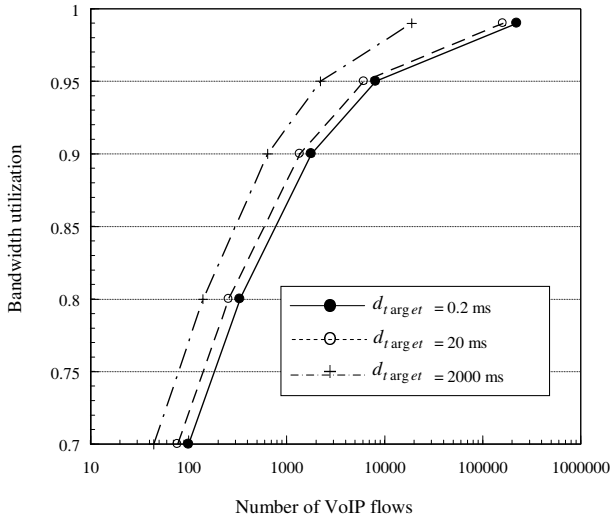

Figure 8: Impact of the delay target d_{target} on the bandwidth utilization: G.723.1

is attained when the number of sources is less than 3000. Note that the bandwidth required for accommodating 3000 IP-telephony connections whose talkspurt activity is 0.5 is about 13 Mbps under 90%-bandwidth utilization. We also show the impact of the delay target d_{target} on the bandwidth utilization when $\epsilon = 10^{-3}$ and $\delta = 0.5$ in Figure 8. The difference in the delay target does not have large impact on the bandwidth utilization. In particular, the difference between the case of $d_{target} = 20$ ms and that of $d_{target} = 0.2$ ms is quite small.

Next, we explain the results for the IP telephony source using G.729 coding. In Figure 9, we show the relationship between the number of multiplexed sources and the bandwidth utilization when $d_{target} = 20$ ms and $\delta = 0.5$. The 90%-bandwidth utilization is also attained when the number of sources is 3000 or less. The impact of the delay target d_{target} on the bandwidth utilization when $\epsilon = 10^{-3}$ and $\delta = 0.5$ is shown in Figure 10, which also indicates that d_{target} does not have large impact on the bandwidth utilization when d_{target} is less than 20 ms.

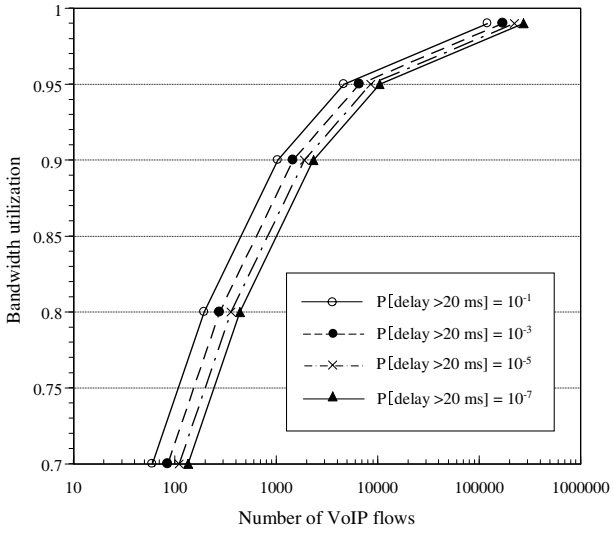

Figure 9: Relationship between the number of multiplexed sources and bandwidth utilization: G.729

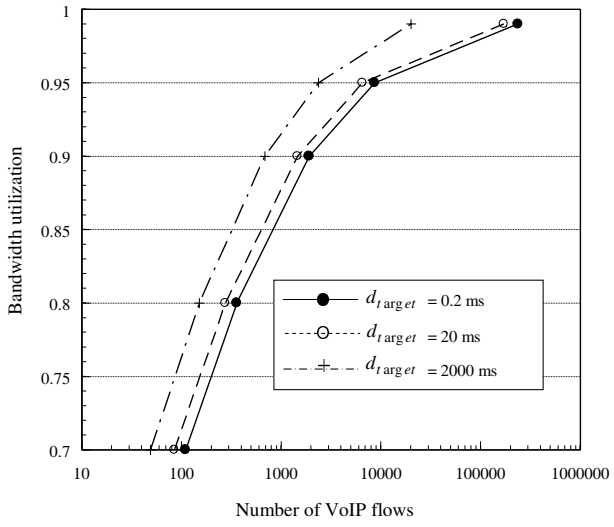

Figure 10: Impact of the delay target d_{target} on the bandwidth utilization: G.729

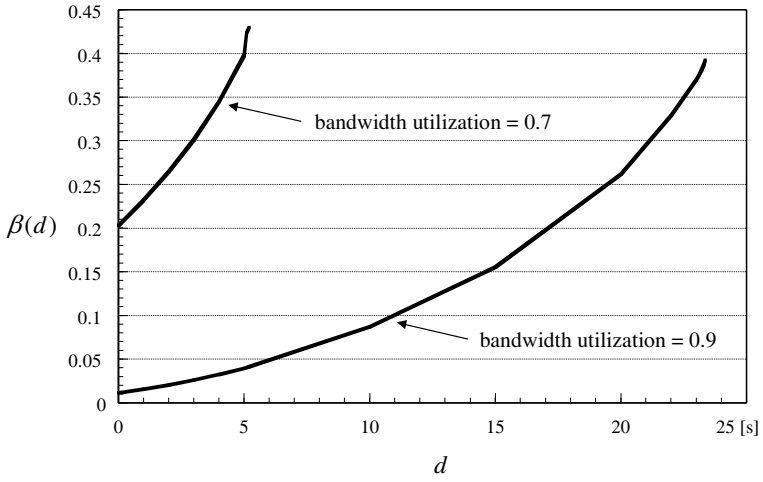

Figure 11: Shape function: G.723.1

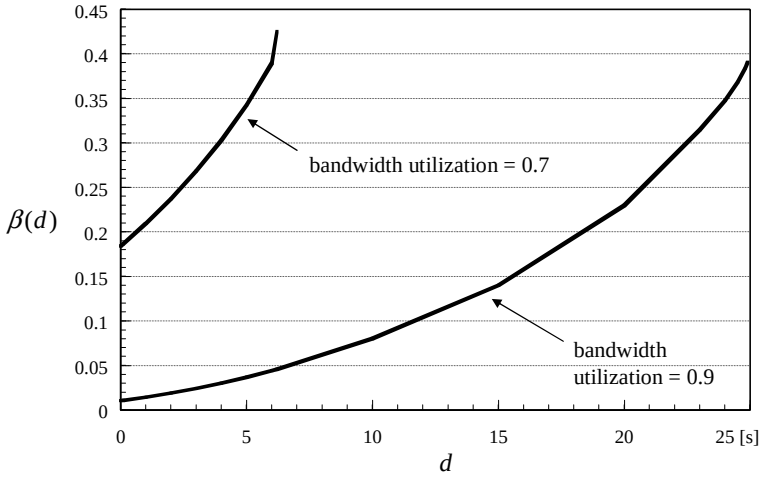

Figure 12: Shape function: G.729

Finally we illustrate the shape function $\eta(d)$ in Figure 11 (G.723.1) and Figure 12 (G.729). We see that $\eta(d)$ is a convex function of d as was explained in Section 4. We have conducted other numerical experiments, which also confirmed that the shape function is convex for other choices of the value of δ.

6. Concluding Remarks

We have derived the upper bounds of the virtual-waiting-time distribution of a single-server queue when several kinds of regulated sources are multiplexed. The statistical multiplexing has been theoretically and numerically analyzed by using the derived upper bound.

Numerical results concerning the shape function reveal that the regulated sources have a noteworthy characteristic in the statistical multiplexing: the regulated sources are more advantageous than Markovian-arrival sources and long-range-dependent sources. This finding is very valuable from the traffic-engineering viewpoint: once sources are smoothed by a regulator having a deterministic subadditive envelope, then they could achieve large statistical multiplexing gain even if original sources have long-range dependence. This could be a counter example to the general belief that any kinds of QoS guarantees will necessitate operating the Internet at very low utilization because of the bursty nature of the Internet traffic. It is also an objection to the argument that a token bucket cannot completely remove the long-range dependence of the traffic [28] and thus it is not useful to enhance the statistical multiplexing gain.

It may be possible to theoretically prove the convexity the shape function although in this paper we have numerically shown it. This remains for further study.

References

[1] R. Agrawal, R. Cruz, C. Okino and R. Rajan: Performance bounds for flow control protocols. *IEEE/ACM Trans. Networking*, **44** (1998), 1096–1107.

[2] F. Baccelli and P. Bremaud: *Elements of Queueing Theory* (Springer-Verlag, 1994).

[3] S. Blake, et al.: An architecture for diffentiated services. *RFC 2475*, (1998).

[4] R. Boorstyn, A. Burchard, J. Liebeherr and C. Oottamakorn: Statistical service assurances for traffic scheduling algorithms. *IEEE J. Select. Areas Commun.*, **18** (2000), 2651–2664.

[5] D. Botvich and N. Duffield: Large deviations, economies of scale, and the shape of the loss curve in large multiplexers. *Queueing Systems*, **20** (1995), 293–320.

[6] R. Braden, D. Clark and S. Shenker: Integrated services in the internet architecture: an overview. *RFC 633*, (1994).

[7] E. Buffet and N. Duffield: Exponential upper bounds via martingales for multiplexers with markovian arrivals. *J. Appl. Prob.*, **31** (1994), 1049–1061.

[8] C. Chang, Y. Chiu and W. Song: On the performance of multiplexeing independent regulated inputs. *ACM SIGMETRICS 2001* (2001), 184–193.

[9] C. Chang, R. Cruz, J. L. Boudec and P. Thiran: A min, + system theory for constrained traffic regulation and dynamic service guarantees. *IEEE/ACM Trans. Networking*, **10** (2002), 805–817.

[10] C. Courcoubetis and R. Weber: Buffer overflow asymptotics for a buffer handling many traffic sources. *J. Appl. Prob.*, **33** (1996), 886–903.

[11] R. Cruz: A calculus for network delay, part I: network elements in isolation. *IEEE Trans. Inform. Theory*, **37** (1991), 114–121.

[12] R. Cruz: A calculus for network delay, part II: network analysis. *IEEE Trans. Inform. Theory*, **37** (1991), 132–141.

[13] N. Duffield: Exponential bounds for queues with markovian arrivals. *Queueing Systems*, **7** (1994), 413–430.

[14] N. Duffield: Economies of scale in queues with sources having power-law large deviation scalings. *J. Appl. Prob.*, **33** (1996), 840–857.

[15] N. Duffield: Conditioned asymptotics for tail probabilities in large multiplexer. *Performance Evaluation*, **31** (1998), 281–300.

[16] A. Elwalid, D. Mitra and R. Wentworth: A new approach for allocating buffers and bandwidth to heterogeneous regulated traffic in an atm node. *IEEE J. Select. Areas Commun.*, **13** (1995), 1115–1127.

[17] K. Iida, T. Takine, H. Sunahara and Y. Oie: Delay analysis for cbr traffic under static-priority scheduling. *IEEE/ACM Trans. Networking*, **9** (2001), 177–185.

[18] V. Jacobson, K. Nichols, et al.: An expedited forwarding PHB. *RFC 2598*, (1999).

[19] G. Kesidis and T. Konstantopoulos: Extremal traffic and worst-case performance for queues with shaped arrivals. *Proc. Workshop Analysis Simulation Commun. Networks* (1998), 9–13.

[20] G. Kesidis and T. Konstantopoulos: Extremal shape-controlled traffic patterns in high-speed networks. *IEEE Trans. Commun.*, **48** (1999), 813–819.

[21] G. Kesidis and T. Konstantopoulos: Worst-case performance of a buffer with independent shaped arrival processes. *IEEE Commun. Letters*, **4** (2000), 26–28.

[22] T. Kostas, M. Borella, I. Sidhu, G. Schuster, J. Grabiec and J. Mahler: Real-time voice-over packet-switched networks. *IEEE Network Mag.*, **12** (1998), 18–27.

[23] A. K. Parekh and R. G. Gallager: A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Trans. Networking*, **1** (1993), 344–357.

[24] F. L. Presti, Z. Zhang, J. Kurose and D. Towsley: Source time scale and optimal buffer/bandwidth tradeoff for heterogeneous regulated traffic in a network mode. *IEEE/ACM Trans. Networking*, **7** (1999), 490–501.

[25] S. Shioda: Worst case performance of ATM multiplexer with GCRA-conforming arrival processes. *Journal of the Operation Research Society of Japan*, **41** (1998), 3–20.

[26] K. Sriram and W. Whitt: Characterizating superposition arrival process in packet multiplexers for voice data. *IEEE J. Select. Areas Commun.*, **4** (1986), 833–846.

[27] D. Stiliadis and A. Varma: Latency-rate servers: a general model for analysis of traffic scheduling algorithms. *IEEE/ACM Trans. Networking*, **6** (1998), 611–624.

[28] S. Vamvakos and V. Anantharam: On the departure process of a leaky bucket system with long-range dependent input traffic. *Queueing Systems*, **28** (1998), 191–214.

[29] M. Vojnović and J. L. Boudec: Bounds for independent regulated inputs multiplexed in a service curve network element. *IEEE Trans. Commun.*, **51** (2003), 735–740.

[30] Y. Yatsuzuka: Highly sensitive speech detector and high-speed voiceband data discriminator in dsi-adpcm systems. *IEEE Trans. Commun.*, **30** (1982), 739–750.

Appendix

A.1. Proof of Theorem 3.1

We begin with proving the following lemma, which will be used in the proof of the theorem.

Lemma A.1. *If* $\sum_{k=1}^{K} N_k \rho^{(k)} t < \beta(t) + \beta(d)$ *for all* t *and* $n \le n_{max}(d, \vec{N})$, *then*

$$P[\sum_{k=1}^{K}\sum_{j=1}^{N_k} A_j^{(k)}(0, nT) > \beta(nT) + \beta(d)]$$
$$\le \prod_{k=1}^{K}\left\{\left(\frac{n\rho^{(k)}T}{\gamma_k(n,T,\theta^*)}\right)^{\gamma_k(n,T,\theta^*)/\alpha^{(k)}(nT)}\left(\frac{\alpha^{(k)}(nT)-n\rho^{(k)}T}{\alpha^{(k)}(nT)-\gamma_k(n,T,\theta^*)}\right)^{1-\gamma_k(n,T,\theta^*)/\alpha^{(k)}(nT)}\right\}^{N_k},$$

where $\theta^*(n, T, d, \vec{N})$ *is the unique solution in* $(0, \infty)$ *to the following equation:*

$$\beta(nT) + \beta(d) = \sum_{k=1}^{K} N_k \gamma_k(n, T, \theta). \tag{A.1}$$

Proof. To prove the existence of $\theta^*(n, T, d, \vec{N})$, observe that $g(n;\theta) \stackrel{\text{def}}{=} \sum_{k=1}^{K} N_k \gamma_k(n, T, \theta)$ is continuous and strictly increasing with θ. In addition to this, if $n \le n_{max}$, then

$$g(n;0) = n\sum_{k=1}^{K} N_k \rho^{(k)} T \le \beta(nT) + \beta(d), \text{ and}$$
$$\lim_{\theta\to\infty} g(n;\theta) = \sum_{k=1}^{K} N_k \alpha^{(k)}(nd) > \beta(nT) + \beta(d),$$

which readily proves the existence of θ^* in $(0, \infty)$ satisfying (A.1). To prove the rest, first note that Chernoff's inequality yields

$$P[\sum_{k=1}^{K}\sum_{j=1}^{N_k} A_j^{(k)}(0, nT) > \beta(nT) + \beta(d)] \le e^{-\theta(\beta(nT)+\beta(d))} E[e^{\theta\sum_{k=1}^{K}\sum_{j=1}^{N_k} A_j^{(k)}(0,nT)}]$$
$$= e^{-\theta(\beta(nT)+\beta(d))} \prod_{k=1}^{K}\prod_{j=1}^{N_k} E[e^{\theta A_j^{(k)}(0,nT)}]$$
$$\le e^{-\theta(\beta(nT)+\beta(d))} \prod_{k=1}^{K}\left(1 + \frac{n\rho^{(k)}T}{\alpha^{(k)}(nT)}(e^{\theta\alpha^{(k)}(nT)} - 1)\right)^{N_k}$$

for all $\theta \geq 0$, where the last inequality follows from Lemma 2.1. Thus,

$$\begin{aligned} & P[\sum_{k=1}^{K}\sum_{j=1}^{N_k} A_j^{(k)}(0,nT) > \beta(nT)+\beta(d)] \\ \leq & \inf_{\theta\geq 0}\left[e^{-\theta(\beta(nT)+\beta(d))}\prod_{k=1}^{K}\left(1+\frac{n\rho^{(k)}T}{\alpha^{(k)}(nT)}(e^{\theta\alpha^{(k)}(nT)}-1)\right)^{N_k}\right]. \end{aligned} \tag{A.2}$$

Next note that

$$\frac{d}{d\theta}\log\left\{e^{-\theta(\beta(nT)+\beta(d))}\prod_{k=1}^{K}\left(1+\frac{n\rho^{(k)}T}{\alpha^{(k)}(nT)}(e^{\theta^*\alpha^{(k)}(nT)}-1)\right)^{N_k}\right\} = -\beta(nT)-\beta(d)+g(n;\theta),$$

which indicates that the infimum of the right hand side of (A.2) is attained when $\theta = \theta^*(n;d,\beta;\vec{N})$ because $g(n;\theta)$ is continuous and strictly increasing with θ. Thus,

$$\begin{aligned} & \inf_{\theta\geq 0}\left[e^{-\theta(\beta(nT)+\beta(d))}\prod_{k=1}^{K}\left(1+\frac{n\rho^{(k)}d}{\alpha^{(k)}(nT)}(e^{\theta\alpha^{(k)}(nT)}-1)\right)^{N_k}\right] \\ &= e^{-\theta^*(\beta(nT)+\beta(d))}\prod_{k=1}^{K}\left(1+\frac{n\rho^{(k)}T}{\alpha^{(k)}(nT)}(e^{\theta^*\alpha^{(k)}(nT)}-1)\right)^{N_k} \\ &= \prod_{k=1}^{K}\left\{e^{-\theta^*\gamma_k(n,T,\theta^*)}\left(1+\frac{n\rho^{(k)}T}{\alpha^{(k)}(nT)}(e^{\theta^*\alpha^{(k)}(nT)}-1)\right)\right\}^{N_k} \\ &= \prod_{k=1}^{K}\left\{\left(\frac{n\rho^{(k)}T}{\gamma_k(n,T,\theta^*)}\frac{\alpha^{(k)}(nT)-\gamma_k(n,T,\theta^*)}{\alpha^{(k)}(nT)-n\rho^{(k)}T}\right)^{\gamma_k(n,T,\theta^*)/\alpha^{(k)}(nT)}\left(\frac{\alpha^{(k)}(nT)-n\rho^{(k)}T}{\alpha^{(k)}(nT)-\gamma_k(n,T,\theta^*)}\right)\right\}^{N_k} \\ &= \prod_{k=1}^{K}\left\{\left(\frac{n\rho^{(k)}T}{\gamma_k(n,T,\theta^*)}\right)^{\gamma_k(n,T,\theta^*)/\alpha^{(k)}(nT)}\left(\frac{\alpha^{(k)}(nT)-n\rho^{(k)}T}{\alpha^{(k)}(nT)-\gamma_k(n,T,\theta^*)}\right)^{1-\gamma_k(n,T,\theta^*)/\alpha^{(k)}(nT)}\right\}^{N_k}. \end{aligned}$$

By substituting the above result into (A.2), we complete the proof. □

Now we are ready to prove the theorem. Define

$$\tilde{M}_n^{(T)}(\vec{N}) \stackrel{\text{def}}{=} \sup_{l;0\leq l\leq n}\{\sum_{k=1}^{K}\sum_{j=1}^{N_k} A_j^{(k)}(-lT,0) - B(-lT,0)\}.$$

Random variable $\tilde{M}_n^{(T)}(\vec{N})$, which is the workload at time 0 when the buffer is empty at time $-nT$, is non-decreasing with n and thus converges to a random variable $\tilde{M}^{(T)}(\vec{N})(\leq\infty)$ as $n\to\infty$. In addition to this, if the stability condition (5) is met, then $\tilde{M}^{(T)}(\vec{N})$ is equal to the steady-state workload in distribution [2, 29]. Observe that

$$\begin{aligned} P[\tilde{M}^{(T)}(\vec{N}) > x] &= P[\sup_{n\geq 1}\{\sum_{k=1}^{K}\sum_{j=1}^{N_k} A_j^{(k)}(-nT,0) - B(-nT,0)\} > x] \\ &= P[\inf\{n\geq 1; \sum_{k=1}^{K}\sum_{j=1}^{N_k} A_j^{(k)}(-nT,0) > B(-nT,0)+x\} < \infty] \end{aligned}$$

$$\begin{aligned} &\leq \sum_{n=1}^{\infty} P[\sum_{k=1}^{K}\sum_{j=1}^{N_k} A_j^{(k)}(-nT,0) > \beta(nT)+x] \\ &= \sum_{n=1}^{n_{max}(T,\vec{N})} P[\sum_{k=1}^{K}\sum_{j=1}^{N_k} A_j^{(k)}(0,nT) > \beta(nT)+x]. \end{aligned}$$

Thus, we have

$$\begin{aligned} P[\tilde{D}^{(T)}(\vec{N}) > d] &\leq P[\tilde{M}^{(T)}(\vec{N}) > \beta(d)] \\ &\leq \sum_{n=1}^{n_{max}(d,\vec{N})} P[\sum_{k=1}^{K}\sum_{j=1}^{N_k} A_j^{(k)}(0,nT) > \beta(nT)+\beta(d)]. \end{aligned} \tag{A.3}$$

The proof is completed by applying Lemma A.1 to (A.3) and by conducting a few calculations.

A.2. Proof of Theorem 3.2

We first prove the following lemma, which will be used in the proof of the theorem.

Lemma A.2. *Let $W(t,\vec{N})$ be the workload (the amount of data) in the buffer at time t in the continuous-time model when the number of multiplexed sources is $\vec{N} \stackrel{\text{def}}{=} (N_1, N_2, \cdots, N_K)$, and $\tilde{W}^{(T)}(t,\vec{N})$ be the workload in the buffer at time t in the discrete-time model when the number of multiplexed sources is $\vec{N}$ and the length of time unit is T. If $B(t,t+\tau)$ has a deterministic upper bound such that*

$$B(t,t+\tau) \leq \hat{\beta}(\tau),$$

then

$$\tilde{W}_n^{(T)}(\vec{N}) \leq W(nT,\vec{N}) \leq \tilde{W}_n^{(T)}(\vec{N}) + \hat{\beta}(T). \tag{A.4}$$

Proof. We prove (A.4) by induction. Notice that $W(nT,\vec{N})$ is given by

$$W(nT,\vec{N}) = \sup_{u;0\leq u\leq t}\{A_{all}(u,t) - B(u,t)\}, \quad A_{all}(u,t) \stackrel{\text{def}}{=} \sum_{j=1}^{N} A_j(u,t).$$

First observe that

$$\begin{aligned} W(T,\vec{N}) &= \sup_{u;0\leq u\leq T}\{A_{all}(u,t) - B(u,t)\} \\ &\geq \max_{n;0\leq n\leq 1}\{A_{all}(nT,T) - B(nT,T)\} = \tilde{W}_1^{(T)}(\vec{N}), \end{aligned}$$

and

$$\begin{aligned} W(T,\vec{N}) &= \sup_{u;0\leq u\leq T}\{A_{all}(u,t) - B(u,t)\} \\ &\leq \max_{n;0\leq n\leq 1}\{A_{all}(nT,T)\} \\ &\leq \max_{n;0\leq n\leq 1}\{A_{all}(nT,T) - B(nT,T) + B(0,T)\} \leq \tilde{W}_1^{(T)}(\vec{N}) + \hat{\beta}(T). \end{aligned}$$

Thus, (A.4) holds when $n = 1$. Now, we assume that (A.4) holds when $n = k$, under which

$$\begin{aligned} W((k+1)T, \vec{N}) &= \max\{W(kT, \vec{N}) + A_{all}(kT, (k+1)T) - B(kT, (k+1)T), \\ &\qquad \sup_{u; kT < u \le (k+1)T} \{A_{all}(u, (k+1)T) - B(u, (k+1)T)\}\} \\ &\ge [W(kT, \vec{N}) + A_{all}(kT, (k+1)T) - B(kT, (k+1)T)]^+ \\ &\ge [\tilde{W}_k^{(T)}(\vec{N}) + A_{all}(kT, (k+1)T) - B(kT, (k+1)T)]^+ \\ &= \tilde{W}_{k+1}^{(T)}(\vec{N}), \end{aligned}$$

and

$$\begin{aligned} & W((k+1)T, \vec{N}) \\ = \; & \max\{W(kT, \vec{N}) + A_{all}(kT, (k+1)T) - B(kT, (k+1)T), \\ & \qquad \sup_{u; kT < u \le (k+1)T} \{A_{all}(u, (k+1)T) - B(u, (k+1)T)\}\} \\ \le \; & \max\{W(kT, \vec{N}) + A_{all}(kT, (k+1)T) - B(kT, (k+1)T), A_{all}(kT, (k+1)T)\} \\ \le \; & \max\{\tilde{W}_k^{(T)}(\vec{N}) + A_{all}(kT, (k+1)T) - B(kT, (k+1)T) + \hat{\beta}(T), A_{all}(kT, (k+1)T)\} \\ \le \; & [\tilde{W}_k^{(T)}(\vec{N}) + A_{all}(kT, (k+1)T) - B(kT, (k+1)T)]^+ + \hat{\beta}(T) = \tilde{W}_{k+1}^{(T)}(\vec{N}) + \hat{\beta}(T). \end{aligned}$$

Thus, (A.4) holds when $n = k + 1$, too, which completes the proof. □

Now we are ready to prove the theorem. We first consider the case where $\hat{\beta}(\tau) = \beta(\tau)$. By letting $n \to \infty$ in (A.4), we obtain $\tilde{W}^{(T)}(\vec{N}) \le_{st} W(\vec{N}) \le_{st} \tilde{W}^{(T)}(\vec{N}) + \beta(T)$ where $W(\vec{N})$ and $\tilde{W}^{(T)}(\vec{N})$ are the steady-state workload in the continuous and discrete time model[§], respectively. Thus,

$$\begin{aligned} P[D(\vec{N}) > d] &\le P[W(\vec{N}) > \beta(d)] \\ &\le P[\tilde{W}^{(T)}(\vec{N}) + \beta(T) > \beta(d)] = P[\tilde{W}^{(T)}(\vec{N}) > \beta(d) - \beta(T)] \\ &\le \sum_{n=1}^{n_{max}(d,\vec{N})} P[\sum_{k=1}^{K} \sum_{j=1}^{N_k} A_j^{(k)}(0, nT) > \beta(nT) + \beta(d) - \beta(T)] \\ &\le \sum_{n=1}^{n_{max}(d,\vec{N})} P[\sum_{k=1}^{K} \sum_{j=1}^{N_k} A_j^{(k)}(0, nT) > \beta((n-1)T) + \beta(d)]. \end{aligned}$$

It is not difficult to see that the last term of the above inequality is equal to the right-hand side of (3.5). Since the virtual waiting time when $\hat{\beta}(\tau) \ge \beta(\tau)$ should be less than that when $\hat{\beta}(\tau) = \beta(\tau)$ in distribution, the desired conclusion follows.

A.3. Proof of Theorem 4.1

Observe that

$$\begin{aligned} & P[D(\vec{N}) > d] \\ \le \; & \sum_{n=1}^{\hat{n}_{max}(T,d,\vec{r})} \prod_{k=1}^{K} \left\{ \left(\frac{x_k(n,T)}{\hat{y}_k(n,T,d,\vec{r})} \right)^{\hat{y}_k(n,T,d,\vec{r})} \left(\frac{1 - x_k(n,T)}{1 - \hat{y}_k(n,T,d,\vec{r})} \right)^{1-\hat{y}_k(n,T,d,\vec{r})} \right\}^{Nr_k} \\ \le \; & \hat{n}_{max}(T,d,\vec{r}) \prod_{k=1}^{K} \left\{ \left(\frac{x_k(n^*,T)}{\hat{y}_k(n^*,T,d,\vec{r})} \right)^{\hat{y}_k(n^*,T,d,\vec{r})} \left(\frac{1 - x_k(n^*,T)}{1 - \hat{y}_k(n^*,T,d,\vec{r})} \right)^{1-\hat{y}_k(n^*,T,d,\vec{r})} \right\}^{Nr_k} \end{aligned}$$

[§]The ordering relation $\le_{st}$ between random variables denotes the strong ordering [2].

$$= \hat{n}_{max}(T, d, \vec{r}) e^{-N\eta(T,d,\vec{r})}, \tag{A.5}$$

which completes the proof. Note that since θ^* exists in $(0, \infty)$ by Lemma A.1,

$$x_k(n^*, T) < \hat{y}_k(n^*, T, d, \vec{r}) \quad (k = 1, \cdots, K).$$

Thus, for all k $(1 \leq k \leq K)$,

$$\left(\frac{x_k(n^*, T)}{\hat{y}_k(n^*, T, d, \vec{r})}\right)^{\hat{y}_k(n^*,T,d;\vec{N})} \left(\frac{1 - x_k(n^*, T)}{1 - \hat{y}_k(n^*, T, d, \vec{r})}\right)^{1-\hat{y}_k(n^*,T,d,\vec{r})} < 1,$$

which guarantees that η is positive.

Shigeo Shioda
Urban Environment Systems
Faculty of Engineering, Chiba University
1-33 Yayoi, Inage, Chiba 263-8522, Japan
E-mail: shioda@faculty.chiba-u.jp