Bandwidth Allocation for Multiple Qualities of Service Using Generalized Processor Sharing

G. de Veciana and George Kesidis, *Member, IEEE*

Abstract— **We consider the asymptotic behavior of the queue length distribution in segregated buffers sharing a deterministic server via a class of generalized processor sharing (GPS) policies. Such policies have been proposed as a means to guarantee individual quality of service constraints to heterogeneous streams in integrated services digital networks. These results exhibit the manner in which spare capacity is shared by statistically multiplexed traffic streams. The framework corresponds to a natural relaxation of a single GPS node subject to (σ, ρ)-constrained flows where, instead of studying the worst case behavior, we consider statistical bounds on the performance of individual traffic streams.**

Index Terms—**Multiservice communication networks, bandwidth allocation, large deviations.**

I. Introduction

ATM-based BISDN networks with heterogeneous applications requiring stringent performance guarantees will need appropriate service provisioning schemes including: buffer/bandwidth allocation, call admission, and call routing [10], [22]. The difficulty of this problem relative to traditional (circuit-switched) telephone networks lies in the multiplexing of multiple types of packetized traffic streams and messages via switches and communication links. In order for streams to share resources, one must guard against traffic fluctuations by inserting buffers. This in turn introduces interactions among relatively bursty traffic streams often rendering performance analysis difficult and the ensuing traffic management schemes inefficient.

To ease the task of managing such a network it is desirable to obtain an equivalent circuit-switched model based on archetypes for different classes of streams, e.g., audio, graphics, low- and high-quality TV, and LAN-to-LAN traffic. For example, suppose a collection of sources, n_j of type $j \in J$, each known to require a *fixed* bandwidth α_j, share a link with capacity c. One can easily check for available spare bandwidth by considering whether

$$\sum_{j \in J} n_j \alpha_j \leq c. \tag{1}$$

Such a scheme, when extended to a network, would resemble traditional telephone systems where connections are set up if physical resources are available to link the source to the destination. Standard techniques used in telephony could then be adapted to packet-switched networks.

When traffic streams are bursty and share a *buffered* link, (1) no longer has a clear-cut counterpart.[1] Indeed, the interactions among traffic streams and resources in such networks is typically not linear or decoupled across the different types of streams. Network designers often resort to tables indicating how many streams of each type can be tolerated while maintaining acceptable performance [21]. Assuming good traffic models are available, the required tables can be obtained by painstaking simulation of switching devices. For networks composed of multiple nodes these simulations can be prohibitively lengthy and inflexible—one needs to simulate the entire network subject to multiple sets of loads to generate the necessary data.

There exists, however, an approximate result for multiple types of streams sharing a *single* large buffer with deterministic service rate c leading to a linear constraint similar to (1). Specifically, suppose a stringent "δ-constraint" of the form

$$\mathbb{P}(W > B) \leq \exp[-\delta B] \text{ (e.g., } \approx 10^{-9}) \tag{2}$$

is to be satisfied by the stationary workload W (or queue length) of the shared buffer of size B; by maintaining such performance constraints, one can limit overflows, loss, or virtual delays in the buffer. For asymptotically large B, one can show that

$$\sum_{j \in J} n_j \alpha_j(\delta) \leq c \Longleftrightarrow \lim_{B \to \infty} \frac{1}{B} \log \mathbb{P}(W > B) \leq -\delta \tag{3}$$

where $\alpha_j(\delta)$, the *effective bandwidth* of stream j, depends on the statistics of the traffic stream and ranges from the mean to the "peak" rate of the traffic type as δ increases [16], [2], [6], [23]. If a stream has slotted arrivals $\{A_n^j\}$, its effective bandwidth is given by

$$\alpha_j(\delta) = \frac{\Lambda_j(\delta)}{\delta}, \Lambda_j(\delta) = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E} \exp[\delta \sum_{i=1}^{n} A_i^j] \tag{4}$$

and can be computed for many of the standard traffic models [16], [2]. The constraint on the right-hand side of (3) does not strictly guarantee that (2) is satisfied; however, it does suggest that for large enough B this might be roughly the case. Assuming such results are valid, call acceptance can be easily carried out by checking if the available capacity is larger than the effective bandwidth of the new stream. Numerous researchers have suggested the use of such results for bandwidth allocation. For early work in this area see [15], [13], [14]. The existence of effective bandwidths for Markov fluid sources was studied via spectral expansions in [12], [11]. General results were obtained via large deviations in [16], [2], [6], [23].

Although this approach is appealing, many further issues need to be addressed before it becomes viable. Our goal herein is to investigate network designs oriented toward *individual* quality of service (QoS) guarantees rather than aggregate performance constraints. Notice that guaranteeing a cell loss probability no worse than 10^{-9} to the aggregate of a *heterogeneous* mix of traffic streams sharing a buffer is not equivalent to making the same performance guarantee to each of the streams individually [9]. If, however, statistically identical streams share a buffer then the aggregate and individual loss constraints are the same.

In order to guarantee *individual* users specific QoS requirements, buffer and bandwidth sharing rules partially isolating their streams from one another must be put into place. A typical approach is to have traffic queue in segregated buffers while sharing the output capacity via service policies roughly corresponding to weighted round-robin, for example, dynamic-time slicing [21] and Generalized Processor Sharing (GPS) [19]—for earlier work see [8]. The GPS concept in conjunction with leaky bucket flow control has proven to be a particularly successful framework in which to guarantee

Manuscript received June 27, 1994; revised July 13, 1995. The research of G. de Veciana was supported in part by the National Science Foundation under Grant NCR-9409722. The research of G. Kesidis was supported by NSERC of Canada.

G. de Veciana is with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA.

G. Kesidis is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., N2L 3G1, Canada.

Publisher Item Identifier S 0018-9448(96)00014-4.

[1] A conservative approach would be to let α_j denote the peak rate of stream j, while an optimistic one would instead assign mean rate. Neither of these approaches is truly useful when traffic is bursty.

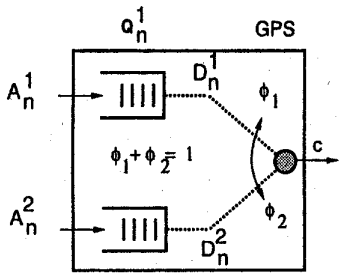

Fig. 1. Segregating traffic and sharing capacity.

individual worst case end-to-end backlog or delay bounds [20]. In this correspondence we focus on a single GPS node subject to statistical traffic flows and individual statistical QoS constraints.

Fig. 1 shows a simple example in which a video stream and a stream corresponding to bulk data transfer share a link with capacity c. The ϕ_1 and ϕ_2 roughly correspond to the fraction of bandwidth allocated to each stream, see [19] and Section II for details. A real-time video stream with stringent delay requirements might be assigned a large ϕ_1, however, when no video packets are queued the whole capacity can be devoted to transmitting data in the second buffer. In the worst case, the GPS policy guarantees a back-logged buffer i a minimal bandwidth $\phi_i c$.

In this correspondence we find rough bandwidth requirements for such systems when each buffer must satisfy an overflow constraint. These in turn indicate how weighting parameters might be selected as well as appropriate call admittance and routing rules based on individual user requirements.

II. Bandwidth Requirements and GPS

Consider a collection of traffic streams, J, sharing multiple segregated buffers, $i = 1 \cdots N$, generalizing the example in Fig. 1. A particular buffer could handle either a single stream, or streams of the same type, e.g., several audio sessions. We will assume that the input traffic streams have known effective bandwidths. The total output rate c is shared among the buffers according to a GPS service policy, see [19], [20] for details. This has the advantage of guaranteeing minimal service rates to each of the buffers according to load-sharing parameters $\{\phi_i\}_{i=1}^N$ preselected by the network software. In particular, the GPS policy guarantees that for any interval of time, say $[-n, 0)$, for which a buffer i has work to do, the cumulative departures from that buffer during the interval, $S_n^{D^i}$, satisfy

$$\frac{S_n^{D^i}}{S_n^{D^j}} \geq \frac{\phi_i}{\phi_j}. \tag{5}$$

This in turn implies that a buffer with a backlog is *guaranteed* a minimum output rate, in particular

$$S_n^{D^i} \geq n \frac{\phi_i}{\sum_{j \in J} \phi_j} c \triangleq n c_i.$$

This fractional guaranteed bandwidth is a worst case estimate. As noted in the Introduction, a buffer i may get a much larger proportion of the total capacity of the link. For example, if all other buffers are idle, buffer i will see a service rate amounting to the total capacity of the output link c. When "spare capacity" is available, it is shared equitably among buffers requiring service, that is, it is shared in proportion to the respective weighting factors.

A. Large Buffer Asymptotics for Two-Buffer GPS

Below we present the case where there are only two buffers sharing a link with capacity c. Without loss of generality, the weighting parameters are normalized (i.e., $\phi_1 + \phi_2 = 1$) and we consider the distribution at Buffer 1. Extensions to multiple streams and buffers are discussed in Section II-B. In our analysis we use discrete-time *fluid* models for traffic streams; these are in turn related to slotted packet streams. Indeed, a consequence of [19, Theorem 2] is that the queue length of a packetized version of GPS (PGPS) also satisfies the asymptotics of Theorem 1 and its corollaries.

Theorem 1: Let $\{A_n^j\}_{n=-\infty}^{\infty}$, $j = 1, 2$ be stationary ergodic arrival processes with $\mathbb{E}[A_n^1 + A_n^2] < c$. Assume the streams are independent and satisfy an LDP (for example, assume they satisfy the Gärtner–Ellis Theorem) with finite asymptotic moment generating functions

$$\Lambda_j(\theta) = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E} \exp[\theta \sum_{i=1}^{n} A_i^j] < \infty \tag{6}$$

with $\Lambda_j^*(\cdot)$ strictly convex. Consider the queue length processes generated when the GPS policy is in effect

$$Q_{n+1}^1 = Q_n^1 + A_n^1 - D_n^1 \text{ and } Q_{n+1}^2 = Q_n^2 + A_n^2 - D_n^2.$$

where the departures, D_n^1 and D_n^2, from the two buffers are consistent with (5). Then the stationary queue length for the first buffer, denoted by the random variable Q^1, satisfies

$$\lim_{B \to \infty} \frac{1}{B} \log \mathbb{P}(Q^1 > B) \leq - \inf_{\alpha_1 + \alpha_2 \wedge \phi_2 c > c} \frac{\Lambda_1^*(\alpha_1) + \Lambda_2^*(\alpha_2)}{\alpha_1 + \alpha_2 \wedge \phi_2 c - c}.$$

Alternatively, for $\delta > 0$

$$\alpha_1(\delta) + \alpha_2^D(\delta, \phi_2 c) \leq c \Rightarrow \lim_{B \to \infty} \frac{1}{B} \log \mathbb{P}(Q^1 > B) \leq -\delta \tag{7}$$

where $\alpha_2^D(\delta, \phi_2 c)$, defined below, depends on bandwidth guaranteed to the second buffer. Indeed,

$$\alpha_2^D(\delta, \phi_2 c) := \begin{cases} \alpha_2(\delta), & \text{if } \alpha_2^*(\theta) < \phi_2 c \text{ and } \mathbb{E}A_n^2 < \phi_2 c \\ \phi_2 c - \Lambda_2^*(\phi_2 c)/\delta, & \text{if } \alpha_2^*(\theta) \geq \phi_2 c \text{ and } \mathbb{E}A_n^2 < \phi_2 c \\ \phi_2 c & \text{if } \mathbb{E}A_n^2 \geq \phi_2 c \end{cases}$$

where $\alpha_2^*(\delta) = \arg\sup_\alpha \{\alpha\delta - \Lambda_2^*(\delta)\}$.

Remark 1: This result has the following intuitive interpretation: In order to guarantee an overflow δ-constraint on Buffer 1 we require that $\alpha_1(\delta) + \alpha_2^D(\delta, \phi_2 c) < c$ (7). In turn, $\alpha_2^D(\delta, \phi_2 c)$ is in one of two regimes. If $\alpha_2^*(\delta) < \phi_2 c$, then $\alpha_2^D(\delta, \phi_2 c) = \alpha_2(\delta)$ so the usual effective bandwidth constraint is obtained, i.e., $\alpha_1(\delta) + \alpha_2(\delta) < c$; otherwise, the *firewall* set up by the GPS service policy comes into play and the requirement depends on the service rate and weighting parameters becoming $\alpha_1(\delta) < \phi_1 c + \Lambda_2^*(\phi_2 c)/\delta$. Eventually if the load on Buffer 2 is high enough, i.e., $\mathbb{E}A_n^2 \geq \phi_2 c$, then the requirement on Buffer 1 becomes $\alpha_1(\delta) < \phi_1 c$. In the first case, we call the system decoupled and say that with respect to the desired constraint it behaves as if the resources were pooled together. In the second case we say that the second buffer is saturated and that segregation is taking place. Further comments on the significance of $\alpha_2^*(\delta)$ are offered in [4]. Buffer 2 will be subject to a similar set of requirements.

Remark 2: We can use previous "devoted"-server effective bandwidth results (see, e.g., [16]) and the minimum bandwidth property of GPS [19] to obtain

$$\lim_{B \to \infty} \frac{1}{B} \log \mathbb{P}(Q^1 > B) \leq - \inf_{\alpha_1 > \phi_1 c} \frac{\Lambda_1^*(\alpha_1)}{\alpha_1 - \phi_1 c}.$$

which is cruder than the equality given by Theorem 2.1.

Proof of Theorem 1: We study the stationary distribution for the queue length in Buffer 1, Q^1, by way of the Loynes's construction [17]. We set the queues empty at time $-m$ and consider the distribution induced at time 0 as the empty starting point $-m$ is moved into the remote past. More specifically, let

$$\begin{aligned} Q_n^{1,m} &= 0, \quad \text{for } n \le -m \\ Q_{n+1}^{1,m} &= Q_n^{1,m} + A_n^1 - D_n^{1,m}, \quad \text{for } n \ge -m \\ &= [Q_n^{1,m} + A_n^1 - c + D_n^{2,m}]^+ \end{aligned}$$

where the last equality uses the fact that GPS is a work-conserving service policy and can be seen to correspond to Lindley dynamics. A similar set of equations defines the dynamics for the second queue, $Q_n^{2,m}$, as well as the dynamics for the aggregate content of the GPS node $Q_n^m = Q_n^{1,m} + Q_n^{2,m}$. Note that the departures $D^{1,m}$ and $D^{2,m}$ have been indexed by m to indicate the fact that the departure processes depend on the starting point $-m$. The stability condition, $\mathbb{E}[A_n^1 + A_n^2] < c$, guarantees the existence of a stationary distribution, denoted by Q, for the total content of the GPS node. Indeed, following Loynes's argument we find that the distribution of Q_0^m increases monotonically to that of $Q = Q^1 + Q^2$; moreover, Q^1 and Q^2 are unique since they can be constructed from the last time the total content of the system emptied before time zero. The stability of the total system guarantees the existence of this time. Thus the sequence $Q_0^{1,m} \le Q_0^m$ in turn converges to Q^1. Define $S_0^{A^1} = 0$ and

$$S_n^{A^1} = \sum_{i=-n}^{-1} A_i^1$$

for $n \ge 1$, and define $S_n^{A^2}, S_n^{D^{1,m}}$ and $S_n^{D^{2,m}}$ similarly. The distribution of $Q_0^{1,m}$ is given by that of a reflected random walk, which in turn has the following form [1, p. 80]:

$$\begin{aligned} Q_0^{1,m} &= \max_{0\le n\le m} S_n^{A^1} - (nc - S_n^{D^{2,m}}) \\ &= \max_{0\le n\le m} S_n^{A^1} - (nc - S_n^{D^{2,m}}) \vee n\phi_1 c \\ &= \max_{0\le n\le m} S_n^{A^1} - (nc - S_n^{D^{2,m}}) \vee (nc - n\phi_2 c) \\ &= \max_{0\le n\le m} S_n^{A^1} + S_n^{D^{2,m}} \wedge n\phi_2 c - nc. \end{aligned} \tag{8}$$

Here $\wedge$ and $\vee$ denote, respectively, min and max. The second equality above follows from the fact that the n achieving the maximum on the right-hand side corresponds to a nonidling busy period, $[-n, 0)$, for which the minimal guaranteed bandwidth property of GPS must hold; whence the additional max term does not affect the right-hand side. The third step uses the fact that $\phi_1 + \phi_2 = 1$.

First assume $\mathbb{E}A_n^2 < \phi_2 c$. Notice that $S_n^{A^1}$ and $S_n^{D^{2,m}}$ in (8) are dependent. This turns out to be inconvenient so we shall replace $S_n^{D^{2,m}}$ with the following upper bound: the conservation of traffic gives

$$S_n^{D^{2,m}} = Q_{-n}^{2,m} + S_n^{A^2} - Q_0^{2,m} \le Q_{-n}^{2,m} + S_n^{A^2}.$$

Since the second queue has a minimal guaranteed bandwidth of $\phi_2 c$, we can upper-bound $Q_{-n}^{2,m}$ by the queue length that would have resulted if in fact the service rate were exactly $\phi_2 c$:

$$Q_{-n}^{2,m} \le \max_{n\le i\le m} S_i^{A^2} - S_n^{A^2} - (i-n)\phi_2 c =: \bar{Q}_{-n}^{2,m}.$$

Combining these two bounds with (8) we obtain the following bound for the queue length in the first buffer:

$$Q_0^{1,m} \le \max_{0\le n\le m} S_n^{A^1} + (\bar{Q}_{-n}^{2,m} + S_n^{A^2}) \wedge n\phi_2 c - nc \tag{9}$$

where $S_n^{A^1}$ and $\bar{Q}_{-n}^{2,m} + S_n^{A^2}$ are now independent.

Since the limits in (6) exist, for $\epsilon > 0$ there is an n_ϵ such that

$$\forall n \ge n_\epsilon, \mathbb{E}\exp[\theta S_n^{A^j}] \le \exp[n(\Lambda_j(\theta) + \epsilon)]$$

for both streams $j = 1$ and 2. We will use this result to study the θ for which $\mathbb{E}\exp[\theta Q_0^{1,m}]$ is finite. Indeed, it follows from (9) and $x, y \ge 0 \Rightarrow \max(x, y) \le x + y$ that

$$\begin{aligned} \mathbb{E}\exp[\theta Q_0^{1,m}] &\le \sum_{n=0}^{m} \mathbb{E}\exp[\theta(S_n^{A^1} + (\bar{Q}_{-n}^{2,m} + S_n^{A^2}) \wedge n\phi_2 c - nc)] \\ &\le C_1 + \sum_{n\ge n_\epsilon} \exp[n(\Lambda_1(\theta) - \theta c + \epsilon)] \\ &\quad \cdot \mathbb{E}\exp[\theta((\bar{Q}_{-n}^{2,\infty} + S_n^{A^2}) \wedge n\phi_2 c)] \end{aligned}$$

where we have used the fact that $\bar{Q}_{-n}^{2,m}$ is nondecreasing in m.

Next we bound the last term on the right-hand side above by showing that following limits exist:

$$\Lambda_2^D(\theta) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\exp[\theta(S_n^{A^2} \wedge n\phi_2 c)] \tag{10}$$

$$= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\exp[\theta((\bar{Q}_{-n}^{2,\infty} + S_n^{A^2}) \wedge n\phi_2 c)] \le \theta\phi_2 c. \tag{11}$$

Note that since the arrival streams satisfy LDP's the limits $\Lambda_2^D(\theta)$ exist, and in fact by Varadhan's lemma [7, p. 120] they are given by

$$\begin{aligned} \Lambda_2^D(\theta) &= \sup_{\alpha_2}[\theta \min[\alpha_2, \phi_2 c] - \Lambda_2^*(\alpha_2)] \\ &= \begin{cases} \Lambda_2(\theta), & \text{if } \alpha_2^*(\theta) < \phi_2 c \\ \theta\phi_2 c - \Lambda_2^*(\phi_2 c), & \text{otherwise} \end{cases} \end{aligned}$$

where $\alpha_2^*(\theta)$ is defined implicitly by

$$\Lambda_2(\theta) = \alpha_2^*(\theta)\theta - \Lambda_2^*(\alpha_2^*(\theta)). \tag{12}$$

This quantity is discussed in [4], where it was called the decoupling bandwidth, an interpretation which also applies in this scenario (see Remark 1).

To show that the limits in (10) and (11) are in fact the same, note that

$$\mathbb{E}\exp[\theta(S_n^{A^2} \wedge n\phi_2 c)] \le \mathbb{E}\exp[\theta((\bar{Q}_{-n}^{2,\infty} + S_n^{A^2}) \wedge n\phi_2 c)]. \tag{13}$$

Also using the fact that

$$\max_i[x_i - y_i] \wedge z = \max_i[(x_i - y_i) \wedge z] = \max_i[x_i \wedge (y_i + z) - y_i]$$

and that the limits in (10) exist we have that for large enough $n > n_\epsilon$,

$$\begin{aligned} &\mathbb{E}\exp[\theta((\bar{Q}_{-n}^{2,\infty} + S_n^{A^2}) \wedge n\phi_2 c)] \\ &\quad = \mathbb{E}\exp[\theta \max_{i\ge n}[S_i^{A^2} - (i-n)\phi_2 c] \wedge n\phi_2 c] \\ &\quad = \mathbb{E}\exp[\theta \max_{i\ge n}[S_i^{A^2} \wedge i\phi_2 c - (i-n)\phi_2 c]] \\ &\quad \le \sum_{i\ge n} \exp[i(\Lambda_2^D(\theta) + \epsilon) - (i-n)\theta\phi_2 c] \\ &\quad \le \sum_{i\ge n} \exp[n(\Lambda_2^D(\theta) + \epsilon) - (i-n)(\theta\phi_2 c - \Lambda_2^D(\theta))] \\ &\quad \le C_2 \exp[(\Lambda_2^D(\theta) + \epsilon)n] \end{aligned} \tag{14}$$

where $C_2 < \infty$ since $\mathbb{E}A_n^2 < \phi_2 c \Rightarrow \Lambda_2^D(\theta) < \theta\phi_2 c$. Thus (13) and (14) sandwich the limit of interest in (11) for large n, which after taking the logarithm and limit as $n \to \infty$ must converge to $\Lambda_2^D(\theta)$ for $\theta > 0$.

Thus we have that

$$\mathbb{E}\exp[\theta Q_0^{1,m}] \le C_1 + C_2 \sum_{n>n_\epsilon} \exp[n(\Lambda_1(\theta) + \Lambda_2^D(\theta) + \epsilon - \theta c)].$$

If $\Lambda_1(\theta) + \Lambda_2^D(\theta) < \theta c$, or equivalently $\alpha_1(\theta) + \alpha_2^D(\theta, \phi_2 c) < c$, then we have that $\mathbb{E}\exp[\theta Q_0^{1,m}] < \infty$. As long as this constraint is in effect it follows by Chebyshev's inequality that for all m, $\mathbb{P}(Q_0^{1,m} > B) \leq C\exp[-\theta B]$; so in fact, letting $m \to \infty$ we have shown that for $\delta > 0$

$$\alpha_1(\delta) + \alpha_2^D(\delta, \phi_2 c) < c \Longrightarrow \limsup_{B\to\infty} \frac{1}{B}\log \mathbb{P}(Q^1 > B) \leq -\delta. \tag{15}$$

On the other hand, if $\mathbb{E}A_n^2 \geq \phi_2 c$, note that by (8)

$$Q_0^{1,m} \leq \max_{0\leq n\leq m}[S_n^{A^1} - n\phi_1 c].$$

A simplification of the above argument yields that $\alpha_1(\delta) < \phi_1 c$ implies the required constraint on Q^1 in (15). □

B. Extension to Multiple Buffers

For a GPS node with $N > 2$ buffers carrying *independent* stationary streams, the tail asymptotics are more complex as there exist a variety of saturation regimes subject to which the spare capacity will be redistributed in a variety of ways according to the parameters $\{\phi_i\}_{i=1}^N$. The following results follow from arguments similar to those in Theorem 1, see [5] for the details.

Corollary 2.1: Let $\{A_n^i\}_{n=-\infty}^{\infty}$, $i = 1, \cdots, N$ be stationary ergodic independent arrival processes with

$$\mathbb{E}\left[\sum_{i=1}^N A_n^i\right] < c$$

satisfying the conditions in Theorem 1. Suppose the streams are segregated into N separate buffers but share a link with capacity c according to a GPS service policy with parameters $\{\phi_i\}_{i=1}^N$, which are positive and normalized, i.e., sum to 1, then

$$\alpha_1(\delta) + \min\left[\sum_{i=2}^N \alpha_i(\delta), (1-\phi_1)c\right] < c$$
$$\Rightarrow \lim_{B\to\infty}\frac{1}{B}\log\mathbb{P}(Q^1 > B) \leq -\delta$$

where Q^1 is the steady-state distribution of the queue length in the buffer associated with the first stream.

III. Summary: Resource Management for Multiple QoS

In [3] and [19], starting from (σ, ρ)-constrained traffic flows, *worst case* bounds were obtained for backlogs and delays in the network. However, the price paid for engineering networks subject to worst case constraints is reduced network "efficiency;" nevertheless, many researchers propose such approaches when dealing with real-time traffic flows requiring multiple qualities of service.

The statistical loss or delay constraints considered herein integrate in a consistent manner with the latter approach. In order to allow multiple QoS, streams are buffered in segregated resources according to their types. Assuming types are statistically identical, the performance constraint on a segregated buffer translates to a per-stream QoS *guarantee*. In order not to lose the gain of statistical multiplexing among types of streams, bandwidth is shared in a work-conserving fashion. The GPS policy gives each buffer a minimal bandwidth guarantee according to network selected weighting factors; these in turn allow control of the overflow characteristics in each buffer.

The results in this correspondence exhibit the interactions among streams in segregated buffers: we have computed approximate bandwidth requirements for a given traffic stream as a function of service weighting parameters and the statistics of the other traffic streams currently sharing the system. This in turn allows for admission control and routing oriented toward multiple qualities of service.

We have concentrated on the single-node scenario when traffic streams satisfy large deviation principles. When traffic streams satisfy "exponential burstiness bounds" (EBB's), stability and upper bounds on queues and delays in a network of GPS nodes have been established subject to appropriate conditions on bandwidth assignment [24], [25]. The relationship between effective bandwidth and (σ, ρ)-bounds was established in [2]; EBB's are a probabilistic relaxation of (σ, ρ) uniform over all time intervals. The effective bandwidth traffic descriptors, considered herein, are specifically chosen to determine "asymptotic" QoS requirements such as the δ-constraint of (3). These require the use of an asymptotic lower bound obtained here via the large deviations principle. We expect that the bandwidth allocation constraints derived here extend to the network setup, subject to stability requirements of the type considered in the previous references, but such results are cumbersome and the simplifications of the previous section may be more usable.

Acknowledgment

The authors wish to thank Zhi-Li Zhang and the anonymous reviewers for many helpful comments and for bringing to their attention an error in a previous version of this correspondence.

References

[1] S. Asmussen, *Applied probability and queues*. Chichester West Sussex, UK: Wiley, 1987.
[2] C.-S. Chang, "Stability, queue length and delay, Part II: Stochastic queueing networks," in *Proc. IEEE CDC* (Tucson, AZ, 1992), pp. 1005–1010.
[3] R. L. Cruz, "A calculus for network delay, Part 1: Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 114–131, 1991.
[4] G. de Veciana, C. Courcoubetis, and J. Walrand, "Decoupling bandwidths for networks: A decomposition approach to resource management for networks," in *IEEE INFOCOM*, vol. 2, 1994, pp. 466–474. Also submitted to *ACM/IEEE Trans. Networking*.
[5] G. de Veciana and G. Kesidis, "Bandwidth allocation for multiple qualities of service using generalized processor sharing," Tech. Rep. SCC-94-01, Univ. Texas Austin, Elec. Comput. Eng. Dep., 1994.
[6] G. de Veciana and J. Walrand, "Effective bandwidths: Call admission, traffic policing and filtering for ATM networks," to appear in *Queueing Systems*, 1995.
[7] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Boston, MA: Jones & Bartlett, 1992.
[8] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," *Internet Res. and Exper.*, vol. 1, 1990.
[9] B. T. Doshi, "Deterministic rule based traffic descriptors for broadband ISDN: Worst case behavior and concept equivalent bandwidth," in *GLOBECOM*, 1993, pp. 1759–1764,
[10] A. E. Eckberg, "B-ISDN/ATM traffic and congestion control," *IEEE Network*, Sept. 1992.
[11] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–343, 1993.
[12] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for multi-type UAS channel," *Queueing Systems*, vol. 9, no. 1, pp. 17–28, 1991.
[13] R. Guérin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks, *IEEE J. Sel. Areas Commun.*, vol. 9, no. 7, pp. 968–981, 1991.
[14] R. Guérin and L. Gun, "A unified approach to bandwidth allocation and access control in fast packet-switched networks," in *IEEE INFOCOM Proc.*, 1992, vol. 1, pp. 1–12.
[15] F. P. Kelly, "Effective bandwidths of multi-class queues," *Queueing Systems*, vol. 9, no. 1, pp. 5–16, 1991.
[16] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, no. 4, pp. 424–428, 1993.
[17] R. M. Loynes, "The stability of a queue with nonindependent interarrivals and service times," *Proc. Camb. Phil. Soc.*, vol. 58, pp. 497–520, 1962.
[18] N. O'Connell, "Large deviations in queueing networks," DIAS Tech. Rep. DIAS-APG-9413, 1994.

[19] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single node case," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 344–357, 1993.
[20] ——, "A generalized processor sharing approach to flow control: The multiple node case," *IEEE/ACM Trans. Networking*, July 1994.
[21] K. Sriram, "Methodologies for bandwidth allocation, transmission scheduling, and congestion avoidance in broadband ATM networks," *Comput. Net. and ISDN Syst.*, vol. 26, pp. 43–59, 1993.
[22] M. Wernik, O. Aboul-Magd, and H. Gilbert, "Traffic management for B-ISDN services," *IEEE Network*, Sept. 1992.
[23] W. Whitt, "Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues," *Telecomm. Syst.*, vol. 2, pp. 71–107, 1993.
[24] O. Yaron and M. Sidi, "Generalized processor sharing networks with exponentially bounded burstiness arrivals," in *IEEE INFOCOM Proc.*, 1994, vol. 2, pp. 628–635.
[25] Z. L. Zhang, D. Towsley, and J. Kurose, "Statistical analyisis of generalized processor sharing scheduling discipline," preprint, 1994.

Lower Bounds on Expected Redundancy for Nonparametric Classes

Bin Yu, *Member, IEEE*

***Abstract*— This correspondence focuses on lower bound results on expected redundancy for universal coding of independent and identically distributed data on [0, 1] from parametric and nonparametric families. After reviewing existing lower bounds, we provide a new proof for minimax lower bounds on expected redundancy over nonparametric density classes. This new proof is based on the calculation of a mutual information quantity, or it utilizes the relationship between redundancy and Shannon capacity. It therefore unifies the minimax redundancy lower bound proofs in the parametric and nonparametric cases.**

I. Introduction

One important ingredient of Rissanen's stochastic complexity theory is his (almost) pointwise lower bound on expected redundancy for regular parametric models, and a minimax counterpart follows from Clarke and Barron [1] (cf. [8]). A similar lower bound was proved by Rissanen *et al.* [11] and Yu and Speed [13] on expected redundancy for the Lipschitz nonparametric class of densities. This lower bound was shown in two different senses: one extending the parametric pointwise bound to an artificial parameter space with a dimension depending on the sample size [11], and the other in the minimax sense [13].

On the other hand, Rissanen's pointwise lower bound can be viewed in the broader picture of the relationship between *redundancy and Shannon capacity*. The study of this useful relationship can be traced back to Gallager [5], who showed that the Shannon capacity serves as a lower bound on the minimax expected redundancy over a parametric source class. Haussler [6] extended the result to general classes of sources. Merhav and Feder [9] showed that the same Shannon capacity is a lower bound on the expected redundancy also in the pointwise or "almost sure" sense. Thus the Shannon capacity serves as a lower bound on the expected redundancy both in minimax and pointwise senses. It follows that in the parametric case the mutual information corresponding to any prior on the parameter space is a lower bound on redundancy in both senses. Using the expansion of the mutual information from a smooth prior in [1], Rissanen's pointwise lower bound can be rederived through this redundancy–capacity paradigm. In general, however, calculating or lower bounding the capacity or mutual information can be difficult.

The focus of this correspondence is on minimax redundancy lower bounds for nonparametric source classes of independent and identically distributed (i.i.d.) data strings. Our contribution is the calculation of the mutual information corresponding to a uniform prior on a specially selected finite source subclass, therefore providing a minimax lower bound on redundancy. Since the old approach for nonparametric minimax lower bounds in [13] is based on accumulated prediction error, not on capacity or mutual information, our current approach unifies the parametric and nonparametric cases.

II. A Review

In this section we review the existing lower bounds on redundancy in the i.i.d. case. For a given i.i.d. data string $x_1, x_2, \cdots, x_n$ and without knowing the distribution f which generated the data, we would like to compress the data in an efficient way. When $f(x) = f_\theta(x)$ belongs to a smooth k-dimensional parametric model class such that the parameter θ can be estimated at the $n^{-1/2}$ rate, Rissanen [10] showed that we need at least $H(f) + \frac{k}{2}\log n$ bits for the string, asymptotically. That is, for any joint density q_n on n-tuples, if we view $-\log q_n(x^n)$ as the code length of an idealized prefix code, then its expected redundancy is

$$E_{f_\theta^n} \log (f_\theta^n / q_n).$$

Rissanen ([10]) showed that

$$\liminf E_{f_\theta^n} \log (f_\theta^n / q_n)/(k \log n/2) \geq 1$$

for all θ values except for a set which depends on q and has Lebesgue measure zero. With a prefix code achieving this lower bound, Rissanen justified that $\frac{k}{2}\log n$ can be viewed as the coding complexity measure of the model class. For more general classes, Merhav and Feder [9] showed that the Shannon capacity replaces $\frac{k}{2}\frac{\log n}{n}$ as the pointwise or almost sure lower bound on redundancy. As we can derive from [1], $\frac{k}{2}\frac{\log n}{n}$ is naturally the leading term in the capacity in the regular parametric case.

When f is known to be in the smooth nonparametric density class of bounded derivatives (or Lipschitz class) on [0, 1], a complexity rate measure of $n^{1/3}$ was established by Rissanen *et al.* in [11] by embedding the nonparametric class in a parametric class of dimension of order $n^{1/3}/\log n$. This embedding reflects the fact that a smooth nonparametric class is in essence a parametric class whose dimension increases with the sample size.

The other approach to obtain lower bounds on expected redundancy is minimax (cf. [2], [3]). Let $w(\theta)$ be a prior on the parameter space and q_n a joint density on n-tuples; then Gallager [5] has shown that

Manuscript received December 4, 1994; revised July 27, 1995. This work was supported by the Army Research Office under Grant DAAH04-94-G-0232 and by the National Science Foundation under Grant DMS-9322817. The material in this correspondence was presented at the IEEE IT/IMS Workshop, Virginia, Oct. 1994.

The author is with the Department of Statistics, University of California, Berkeley, CA 94720-3860 USA.

Publisher Item Identifier S 0018-9448(96)00009-0.