Title
Multiplexing gains in bit stream multiplexors

Permalink
https://escholarship.org/uc/item/99w4m9p2

Journal
IEEE/ACM Transactions on Networking, 3(6)

Authors
Sidhu, Ikhlaq
Jordan, Scott

Publication Date
1995-12-01

DOI
10.1109/90.477724

Peer reviewed

Multiplexing Gains in Bit Stream Multiplexors

Ikhlaq Sidhu, *Member, IEEE,* and Scott Jordan, *Member, IEEE*

***Abstract*—We are concerned with characterizing the variation of multiplexing gains with source type and burstiness in integrated service systems such as ATM. We model a fixed capacity high speed bit pipe that multiplexes a moderate number of bit streams with minimal buffer under a low loss constraint. Each service type is defined by its instantaneous bitrate distribution, but the bitrate distribution of multiplexed streams is approximated as Gaussian. The Gaussian approximation is not as accurate as Chernoff bounds, but it allows for stronger characterization of multiplexing gains. We consider three schemes for allocating bandwidth to services: by individual user, by path and service type, and by path only. We find explicit formulae for sensitivities of required capacity to source rate mean and variance and to loss rate. We characterize multiplexing gains and costs to identify the benefits of each allocation policy. We find that the capacity savings resulting from sharing resources is proportional to the square root of the ratio of source rate variance to source rate mean. This suggests that although bursty sources require more bandwidth, multiplexing gains are increasing with burstiness. We also find that the extra capacity required to multiplex dissimilar source types is increasing with the difference between their burstinesses. This suggests that when bit streams are partially grouped, it is most important first, to group similar source types.**

I. Introduction

WE ARE currently witnessing unprecedented mergers in the telecommunications industry. Regulatory barriers between market segments are falling and major players in the market are positioning themselves to be able to offer as many services as possible. They perceive many communications services are complementary in nature, and can be more efficiently offered over a single system than over separate dedicated facilities. This coalescence of telecommunication services requires new capabilities of the underlying networks that will provide these services. Historically, voice services have required guaranteed performance at a continuous bitrate. Data services, on the other hand, have had variable bitrates but no guaranteed performance. Integration of these services onto a common platform requires that the network be able to offer variable bitrate services with guaranteed performance. This capability is also expected to be used by new services, such as, compressed video.

The problem of guaranteeing performance for differing, bursty services is multifaceted. Recent work on source modeling includes on–off sources, fluid models, and effective bandwidth (see e.g., [1], [3], [4]). Recent work on source policing includes leaky bucket flow control (see e.g., [2], [10], [14], [16]). We are interested here, in understanding when resources should be pooled and for what types of sources. Services are to be multiplexed onto common channels with the expectation that resource sharing will result in a more effective use of resources. However, it is still unclear as to which service types are more suitable for resource multiplexing. Although we do expect multiplexing gains to increase with greater resource sharing, we must keep in mind that the complexity of the sharing scheme is also increasing and the benefits of resource sharing should outweigh the cost of additional complexity.

In ATM, capacity will be allocated to virtual paths on which a number of virtual circuits will be multiplexed. It has not yet been decided how to group virtual circuits, and how to allocate bandwidth to virtual paths in order to guarantee the quality of service of each connection. The level of resource sharing, and resulting multiplexing gain, depend on what service types will share resources. Circuit switching represents *no sharing*. If virtual paths contain virtual circuits with identical paths and source types, here called *homogeneous sharing*, then a traditional statistical multiplexing gain is achieved. If virtual paths contain virtual circuits with *different* source types, here called *complete sharing*, then a further gain will be achieved. These gains must be gauged and compared to each method's complexity.

Previous studies have identified two distinct components of multiplexing related congestion; cell scale congestion and burst scale congestion [13], [15]. Cell scale congestion is caused by cells from many independent sources simultaneously arriving at a node for outbound transmission on the same link. Cell scale congestion can occur even when the average incoming source rates are substantially less than the available capacity on the virtual path connection. Buffers required to avoid cell scale congestion are relatively small, typically on the order of the number of services feeding the virtual path connection. It has also been shown that buffer delay due to cell scale congestion is small compared to the interarrival times for the particular sources. Burst scale congestion, on the other hand, is caused by high speed bursts which may temporarily overload the virtual channel capacity. Buffer lengths must be quite large to avoid this type of congestion and significant queuing delays will be observed. Resource allocation proposals, therefore, generally either allocate capacity close to the peak rate and use short buffers to avoid cell scale congestion, or allocate capacity close to the mean rate and use long buffers to avoid burst scale congestion. The former approach results in low channel utilization for bursty traffic, while the latter approach results in higher channel utilization, higher traffic delay, and substantially larger buffers.

Manuscript received March 7, 1994; revised April 25, 1995; approved by IEEE/ACM Transactions on Networking Editor D. Mitra. This work was supported by the National Science Foundation under Grant NCR-9109223.

I. Sidhu is with the Corporate Systems Division, US Robotics, Skokie, IL 60076-2999 USA.

S. Jordan is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208-3118 USA.

IEEE Log Number 9415908.

1063–6692/95$04.00 © 1995 IEEE

One approach that tends toward the second option is *effective bandwidth.* In the buffered case [4]–[6], [11], each source type is assigned an effective bandwidth, and a new call is admitted if the sum of effective bandwidths for all multiplexed sources remains less than the channel capacity. The multiplexing gain, combined with buffering, successfully lets the system avoid burst level congestion. Sources must have a common loss probability, and effective bandwidth is accurate for small loss probabilities and large buffers.

In contrast, some prefer the simpler approach of allocating a higher capacity and reducing buffer length to the minimum. This approach results in simpler traffic descriptors, simpler policing controls, and reduction in delay times [15]. Commonly, a source's type is defined by its instantaneous distribution of bitrate, and each source requires a minimum loss rate. For a given total bandwidth C, the *acceptance region* is defined as the set of all combinations of numbers of each service type that can be simultaneously accommodated while meeting the minimum loss requirements. The source types are often thought of as different services provided by the network, such as video, voice and data. Several efforts have been made to determine the acceptance region. Hui initially suggested that the bitrate distribution of the sum of several sources could be approximated by a Gaussian distribution [7]. He noted that this approximation is not accurate in the tail, and thus provides inaccurate information for systems with low loss rates. Hui then derived a more accurate bound on the acceptance region using a Chernoff bound and large deviations theory [7], [8]. Hui [7] and Kelly [9] further suggest linear approximations to the acceptance region boundary that are often accurate. Mitra and Morrison [12] derive a uniform asymptotic approximation (UAA) for on–off sources with heterogenous channel requirements, and use this to characterize the acceptance region.

While the Chernoff bound and UAA produce more accurate approximations to the acceptance region than the Gaussian approximation, it is difficult to use them to find sensitivities of system capacity to source parameters, or to find the variation of multiplexing gain with source burstiness. Since these are our goals, we adopt the rougher Gaussian approximation. Our results are, therefore, not strictly valid for non-Gaussian sources at low loss probabilities. The analysis is also restricted to sources with common overflow probability requirements, and systems with no buffer.

In this paper, we consider the capacity required to multiplex a given set of sources under a maximum overflow criterion under complete sharing (capacity C_{CS}); under homogenous sharing (C_{HS}); and under no sharing (C_{NS}). We define two types of multiplexing gain: *complete sharing gain* ($C_{HS} - C_{CS}$) and *homogenous sharing gain* ($C_{NS} - C_{HS}$). We define a source's burstiness as the ratio of its bitrate variance to mean. We show the complete sharing gain, defined as the extra capacity required when only like type services are sharing resources over that required when all services share, is increasing with the square root of burstiness. We also show that the homogenous sharing gain, defined as the additional capacity required under no sharing over that required for homogenous sharing, is bounded by a quantity that also increases with the square root of burstiness. These results suggest resources are more effectively used when separate allocations are combined. Indeed, although bursty sources require more bandwidth, multiplexing gains are increasing with burstiness. Finally, we define a *diversity cost* that measures the excess capacity required by a system with two services using complete sharing over one in which all capacity is devoted to only one service type. We find that this diversity cost increases with the difference in the burstinesses of the two source types. This result suggests that when bit streams are partially grouped, it is most important to first group similar source types.

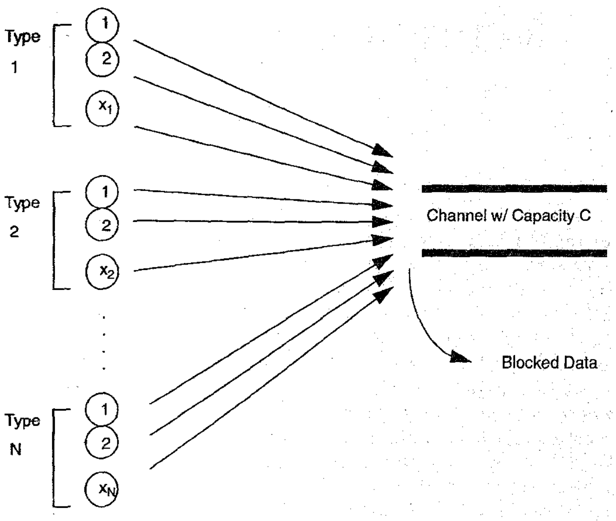


Fig. 1. Bit stream multiplexing block diagram.

Our model is presented in Section II. The acceptance region under a complete sharing discipline is introduced in Section III. We derive approximations to this region in Sections IV and V. In Section VI, we derive the sensitivity of required capacity to blocking probability, and source rate mean and variance. Finally in Section VII, we show that sharing of resources can obtain the same blocking probability as partitioning of resources at a capacity savings proportional to the square root of the ratio of the source rate variance to source rate mean.

II. Bit Stream Multiplexing Model

In this section we define a bit stream multiplexing model and three resource allocation disciplines. The model addresses N different types of bit stream generating sources multiplexed onto a fixed capacity channel. A schematic of the model is shown in Fig. 1.

Bit stream sources generate data (in bits or cells) at a time varying rate. The combined data stream shares a channel of capacity C in accordance with a specific multiplexing discipline. There is no buffer in this model. We will evaluate the effectiveness for the following three alternative multiplexing disciplines: *complete sharing*, *homogeneous sharing*, and *no sharing*.

Sources are classified as one of N different source types. Each source is defined by an instantaneous data rate distribution. Sources of the same source type are indexed. We assume each source has a data rate that is an ergodic random process and that the sources are independent of each other. We adopt the following notation:

N	Number of source types.
S	Set of source types, $S = \{1, 2, \cdots, N\}$.
$x(s)$ or x_s	Number of sources of type $s \in S$, $x(s) \geq 0$. Type s sources are indexed $(1, 2, \cdots, x_s)$.
X	Operating point, $X = [x(1), x(2), \cdots, x(N)]$.
$r_{s,i}(t)$	Time varying data rate for the ith source of type $s \in S$, $i \in \{1 \ \cdots \ x(s)\}$.
$f_s(r)$	Density function of the instantaneous distribution of data rate $r_{s,i}(t)$ of source type $s \in S$.
C	Capacity of the channel.
C_{CS}	Capacity required to multiplex using the *complete sharing* discipline.
C_{HS}	Capacity required to multiplex using the *homogeneous sharing* discipline.
C_{NS}	Capacity required to multiplex using the *no sharing* discipline.
$PB_{\max}$	Worst case acceptable level of PB, indicating the desired quality of service.
A	*Acceptance region*, the set of operating points X resulting in an acceptable quality of service.

Four examples of source types are defined in Fig. 2 for future use. The density function $f_s(r)$ and a possible sample path of $r_{s,i}(t)$ are shown for each. We note that the lack of a buffer at the multiplexing point implies that the marginal density of each source is sufficient to specify the distribution of the rate of the combined data stream. No joint densities are needed. In accordance with an allocation discipline, data streams from all sources share a channel with fixed capacity C. We define blocking to be the event that some portion of the combined data stream is not accepted by the channel. In this paper, we assume all sources have identical maximum loss requirements and that any loss is split among the constituent source streams proportional to the mean rate.

We present three alternative disciplines to allocate the channel between sources. These disciplines are presented in decreasing order of resource sharing or multiplexing; complete sharing, homogeneous sharing, and no sharing.

In complete sharing, all source types share the entire channel. Blocking occurs only when the combined data stream is greater than the channel capacity $C = C_{CS}$

$$B_{CS} = \left[\sum_{s \in S} \sum_{j=1}^{x(s)} r_{s,j}(t) > C_{CS} \right]. \tag{1}$$

In homogeneous sharing, the channel of capacity C is partitioned into $C_{HS,1}, C_{HS,2}, \cdots, C_{HS,N}$ and each subchannel is devoted to its respective source type. For example, a 1 type source cannot use capacity $C_{HS,2}$ even if $C_{HS,1}$ is fully utilized and $C_{HS,2}$ is free. A particular homogeneous sharing policy is defined by a choice of partition

$$C = \sum_{s \in S} C_{HS,s}. \tag{2}$$

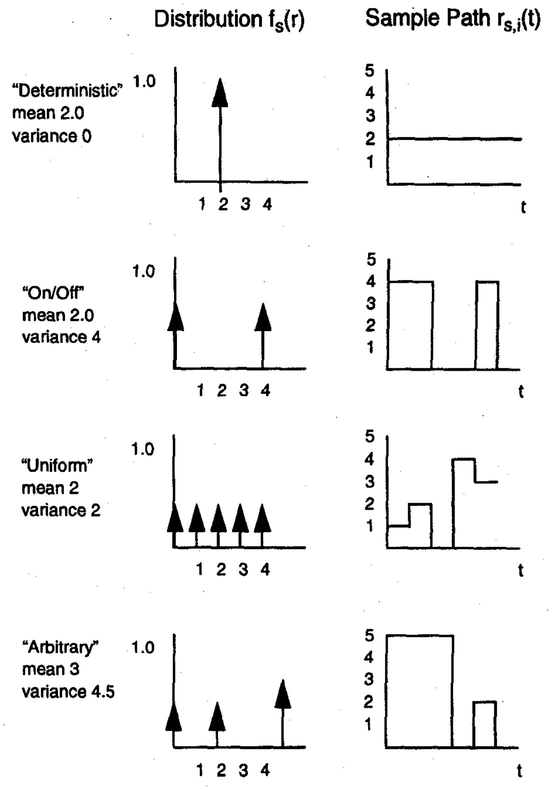


Fig. 2. Examples of bit stream generating sources.

Blocking occurs when a partition is overloaded

$$B_{HS,s} = \left[\sum_{j=1}^{x(s)} r_{s,j}(t) > C_{HS,s} \right], \quad s \in S. \tag{3}$$

In the no sharing discipline, there is no sharing of capacity between sources. Each source of type $s \in S$ is allocated a capacity of $C^{o}_{NS,s}$. A particular no sharing policy is defined by a choice of partition

$$C = \sum_{s \in S} x(s) C^{o}_{NS,s}. \tag{4}$$

Blocking occurs if an individual bit stream generating source exceeds it's allocated capacity

$$B_{NS,s,j} = [r_{s,j}(t) > C^{o}_{NS,s}], \qquad s \in S, \, j \in 1, \cdots, x(s). \tag{5}$$

Given a channel of capacity C, we define the *complete sharing acceptance region* as the set of all operating points $X = [x(1), x(2), \cdots, x(N)]$ that satisfy the service quality criterion that $P(B_{CS}) \leq PB_{\max}$. Similarly, for each homogeneous sharing partition, we can find the set of all operating points X that satisfy $P(B_{HS,s}) \leq PB_{\max} \forall s$. We define the homogeneous sharing acceptance region as the *union of such sets over all possible partitions.* The *no sharing acceptance region* is similarly defined as the union (over all partitions) of sets of operating points satisfying $P(B_{NS,s,j}) \leq PB_{\max} \ \forall s, j$. We further discuss the relationship between blocking probability and loss in future sections.

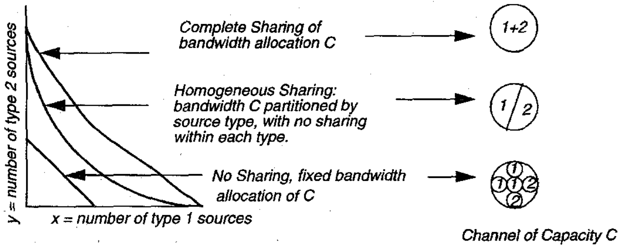

Complete Sharing of bandwidth allocation C
1+2
Homogeneous Sharing: bandwidth C partitioned by source type, with no sharing within each type.
1
2
No Sharing, fixed bandwidth allocation of C
1
1
1
2
2
y = number of type 2 sources
x = number of type 1 sources
Channel of Capacity C

Fig. 3. Acceptance regions for each discipline.

Fig. 3 depicts these acceptance regions for the case of $N = 2$ source types. Since the channel of capacity C can be considered to be a resource, we call the upper boundary of each acceptance region a *resource allocation constraint*. Each acceptance region is the set of first quadrant points contained within the two axis and its corresponding constraint.

The x-axis gives the number of type 1 sources and the y-axis gives the number of type 2 sources that may be multiplexed together. Points above and/or right of each constraint curve do not satisfy the constraint on blocking probability and thus do not provide the acceptable level of service under the corresponding discipline for any partition. The point on each curve with $x = 0$ indicates the maximum number of type 2 sources that can be accommodated by the channel with the prescribed quality of service, and is achieved in homogeneous and no sharing disciplines when all capacity is allocated to type 2 services. By reducing the number of type 2 sources, more type 1 sources can be accommodated. The complete sharing constraint is shown farthest from the origin. For homogenous sharing, a choice of $C_{HS,1}$ and $C_{HS,2}$ results in a rectangular acceptance region. The union of all regions such that $C_{HS,1} + C_{HS,2} = C$ results in the homogenous sharing curve shown above. The no sharing discipline is the same as fixed bandwidth allocation. Disciplines that allow more sharing can accommodate more services. In the next three sections, we investigate the form of the complete sharing resource allocation constraint. In Section VII, we return to consideration of the homogeneous and no sharing resource allocation constraints.

III. Acceptance Region for Complete Sharing

In this section, we present the resource allocation constraint under a complete sharing discipline. We illustrate this constraint for three sample cases using bit stream generating sources from Fig. 2. The example cases have been constructed with $N = 2$ source types. We will build on these examples in later sections.

The combined data rate $R(t)$ at a fixed time t and an operating point X is the sum of the rates of each source

$$R(X, t) = \sum_{s \in S} \sum_{j=1}^{x(s)} r_{s,j}(t). \tag{6}$$

Since these rates are independent random variables, the distribution $f_R(r)$ of $R(t)$ can be expressed by the following convolution equation:

$$f_{R(X)}(r) = f_1(r)^{(x(1))} * f_2(r)^{(x(2))} * \cdots * f_N(r)^{(x(N))} \tag{7}$$

where $f^{(x)}$ denotes the x-fold convolution of f.

Given a specific channel capacity C and maximum blocking probability $PB_{\max}$, the complete sharing constraint $P(B_{CS}) \leq PB_{\max}$ is satisfied if and only if the operating point is in the *complete sharing acceptance region* given by

$$\{X \mid F_{R(X)}(C) \geq 1 - PB_{\max}\} \tag{8}$$

where F_R is the distribution function corresponding to the density function in (7).

The corresponding probability of cell loss (9) is given by dividing the average overflow rate by the mean rate of the offered stream. The reader may refer to [7] for a model relating cell loss to burst lengths

$$PL = \frac{\int_C^{\infty} [r - C] f_{R(X)}(r)\, dr}{\int_0^{\infty} r f_{R(X)}(r)\, dr}. \tag{9}$$

As an illustration, we select parameters $C = 100$ and $PB_{\max} = 10^{-3}$. Using the examples from Fig. 2, the constraint was numerically evaluated for the following three cases of source type pairs: i) deterministic versus on–off, ii) on–off versus uniform, and iii) uniform versus arbitrary. The result is shown in Fig. 4. The points shown correspond to the complete sharing boundary.

We note that for the deterministic data rate type source, the number of allowable sources is the channel capacity divided by the data rate, here 50. A source type with the same mean but larger variance (e.g., uniform or on–off sources in Fig. 2) is more bursty and, as we expect, fewer of the more bursty sources can be accommodated. In addition, we observe that each boundary is nearly linear. This leads us to explore a few continuous approximations in the next section.

IV. Gaussian Approximation to Complete Sharing Acceptance Region

Although the complete sharing resource allocation constraint can be obtained using convolutions as shown in the last section, the method is calculation intensive, and furthermore, a closed form result is not available. In this and the next section, we analyze two approximations of the complete sharing boundary suggested by Hui [7]. Each approximation can be written in closed form. In addition, using these approximations, we bound the extra capacity required to multiplex two source types over that required if the entire capacity is used for one source type.

Fix the operating point X. Suppose that the source rates means and variances are μ_s and σ_s^2, respectively, for source types $s \in S$. The instantaneous mean and variance of the combined data stream rate $R(X, t)$ (6) are then given by

$$\begin{aligned} \mu_{R(X)} &= \sum_{s \in S} \mu_s \cdot x_s \\ \sigma_{R(X)}^2 &= \sum_{s \in S} \sigma_s^2 \cdot x_s. \end{aligned} \tag{10}$$

Since the distribution $f_{R(X)}(r)$ is obtained through convolutions of $f_s(r)$, $s \in S$ (7), by the central limit theorem, $f_{R(X)}(r)$ approaches a Gaussian distribution as the number of sources $x(s)$ of each type increases. This results in the first approximation.

Gaussian approximation:

$$R(X,t) \sim N[\mu_{R(X)}, \sigma_{R(X)}]. \tag{11}$$

Given a specific channel capacity C and maximum blocking probability $PB_{\max}$, the corresponding approximate complete sharing acceptance region is given by

$$\left\{ x \,\middle|\, \Phi\left(\frac{C-\mu_{R(X)}}{\sigma_{R(X)}}\right) \le 1 - PB_{\max} \right\}. \tag{12}$$

This can be written more easily using (10) as those operating points X satisfying

$$\sum_{s=1}^{N} \mu_s \cdot x_s + \rho \sqrt{\sum_{s=1}^{N} \sigma_s^2 \cdot x_s} \le C \tag{13}$$

where

$$\rho = \Phi^{-1}(1 - PB_{\max}).$$

Call the upper boundary of this subset the *Gaussian constraint surface*. Note that this surface is entirely specified by the source rate means and variances, the channel capacity C, and the service quality $PB_{\max}$ or ρ. The entire distribution for each source is not required. If the combined rate distribution is not taken as Gaussian, then the number of standard deviations C must be above the mean is $\rho(X)$ where X is the vector of numbers of sources being multiplexed. We consider this more general case in the appendix.

For a solution of (13) (at equality) for x_i in terms of x_j $i \neq j \in S$, see (14) at the bottom of the page. A second solution can be found to have a similar form but it provides inappropriate values for positive valued square roots in (13) and is disregarded. Equations (13) and (14) simplify for the two source case ($N = 2$), see (15), (16) at the bottom of the page. In Fig. 5, we overlay the Gaussian approximation to the complete sharing acceptance region (curve) with the exact region resource allocation constraint (dots) for the three examples of the previous section.

The Gaussian approximation fits the complete sharing acceptance curves well in these examples. The error is always at least $O(1)$, since the exact curve is discrete and the approximation is continuous, however, in these examples the error is usually less than one. The approximation does introduce higher errors when there are few sources with significant rate variance, e.g., near the lower right in Fig. 5(a). In addition, the approximation may not fit as well at lower blocking probabilities, depending on the form of the tail of the source rate distribution [7]. Our primary use of the approximation, however, will be to gauge multiplexing gains.

For the Gaussian approximation, the mapping between blocking probability and loss probability (9) becomes

$$PL = \frac{\sigma_{R(X)}}{\mu_{R(X)}} \int_{\rho}^{\infty} [z - \rho]\varphi(z)\, dz \tag{17}$$

where $\varphi(\cdot)$ is the density of the standard normal distribution.

Previous studies have shown that a complete sharing acceptance region based on the Chernoff bound has a convex complement in the first quadrant [7], [9]. We find that the Gaussian approximate constraint surface (13) also has a convex complement in the first quadrant when $\mu_{R(X)} < C$, i.e., $PB_{\max} < 0.5$.

Theorem 1: The Gaussian approximate acceptance region has a convex complement in the first quadrant if $PB_{\max} < 0.5$.

Proof: Define

$$C(X) = \sum_{s=1}^{N} \mu_s \cdot x_s + \rho \sqrt{\sum_{s=1}^{N} \sigma_s^2 \cdot x_s} \tag{18}$$

so that (13) now becomes $C(X) \le C$.

We first show that $C(X)$, $X \in R_{+}^{N}$, is a concave function of X.

Consider two arbitrary points X^1 and $X^2 \in R_{+}^{N}$. Substituting in (13), we get

$$C(X^1) = \sum_{s=1}^{N} \mu_s \cdot x_s^1 + \rho \sqrt{\sum_{s=1}^{N} \sigma_s^2 \cdot x_s^1}$$

$$C(X^2) = \sum_{s=1}^{N} \mu_s \cdot x_s^2 + \rho \sqrt{\sum_{s=1}^{N} \sigma_s^2 \cdot x_s^2} \tag{19}$$

$$x_i = \frac{\frac{\rho^2\sigma_i^2}{\mu_i} + 2\left(C - \sum_{j\neq i} \mu_j x_j\right) - \rho \sqrt{\left[\frac{\rho\sigma_i^2}{\mu_i} + 2\left(\frac{C - \sum_{j\neq i}\mu_j x_j}{\rho}\right)\right]^2 - 4\left[\left(\frac{C - \sum_{j\neq i}\mu_j x_j}{\rho}\right)^2 - \sum_{j\neq i}\sigma_j^2 x_j\right]}}{2\mu_i} \tag{14}$$

$$C = \mu_1 \cdot x_1 + \mu_2 \cdot x_2 + \rho\sqrt{\sigma_1^2 \cdot x_1 + \sigma_2^2 \cdot x_2} \tag{15}$$

$$x_2 = \frac{\frac{\rho^2\sigma_2^2}{\mu_2} + 2(C - \mu_1 x_1) - \rho\sqrt{\left(\frac{\rho\sigma_2^2}{\mu_2}\right)^2 + \frac{4\sigma_2^2 C}{\mu_2} + 4x_1\left(\sigma_1^2 - \sigma_2^2\frac{\mu_1}{\mu_2}\right)}}{2\mu_2} \tag{16}$$

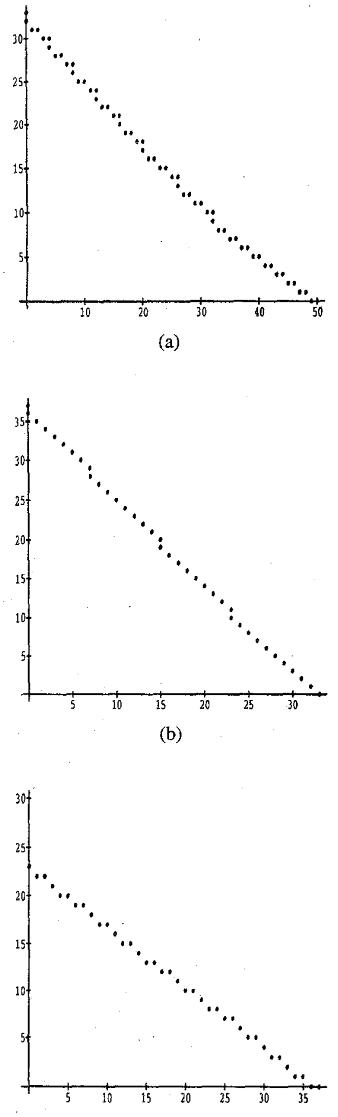


Fig. 4. Acceptance regions under a complete sharing discipline for three example source pairs: (a) Type 1: deterministic, Type 2: on–off; (b) Type 1: on–off, Type 2: uniform; (c) Type 1: uniform, Type 2: arbitrary.

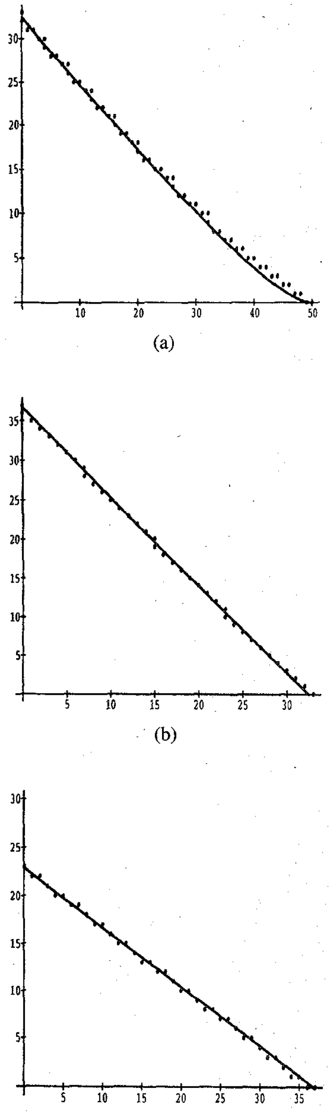


Fig. 5. Gaussian approximation to complete sharing acceptance region for three examples: (a) Type 1: deterministic, Type 2: on–off; (b) Type 1: on–off, Type 2: uniform; (c) Type 1: uniform, Type 2: arbitrary.

$C(X)$ is concave iff

$$C[\lambda X^1 + (1-\lambda)X^2] \ge \lambda C(X^1) + (1-\lambda)C(X^2) \quad \text{for} \quad \lambda \in [0, 1].$$

$PB_{\max} < 0.5$ implies that $\rho > 0$, therefore using (13) and (19), this occurs iff

$$\sqrt{\lambda\left(\sum_{s=1}^{N}\sigma_s^2\cdot x_s^1\right)+(1-\lambda)\left(\sum_{s=1}^{N}\sigma_s^2\cdot x_s^2\right)} \ge \lambda\sqrt{\sum_{s=1}^{N}\sigma_s^2\cdot x_s^1}+(1-\lambda)\sqrt{\sum_{s=1}^{N}\sigma_s^2\cdot x_s^2}.$$

However

$$\sqrt{\lambda a+(1-\lambda)b}\ge\lambda\sqrt{a}+(1-\lambda)\sqrt{b}$$

for any positive constants a, b, and for $\lambda \in [0, 1]$. It follows that $C(X)$ is a concave function of X on $X \in R_+^N$.

Define $\overline{A}(C) = \{X \mid C(X) > C\}$, the first quadrant complement of the Gaussian approximate acceptance region. $\overline{A}(C)$ is an upper level set of $C(X)$, and every upper level set of a concave function is convex. It follows that $\overline{A}(C)$ is a convex set, and the theorem is proved.

V. Linear Approximation to Complete Sharing Acceptance Region and Diversity Cost

The near linearity of the exact acceptance constraint in Fig. 4 leads us to explore a linear approximation. We will derive such an expression by further approximating the Gaussian constraint. The Gaussian approximate resource allocation constraint is a $N - 1$ dimensional surface. Consider a linear

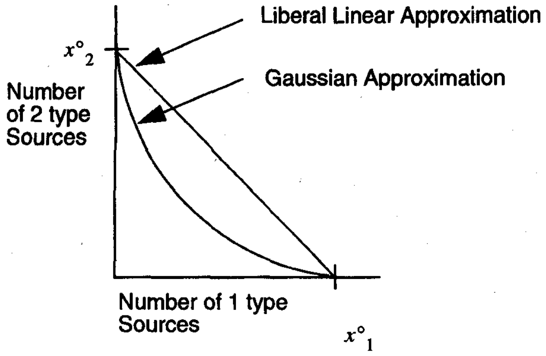


Fig. 6. Liberal linear approximation of resource allocation constraint curve.

approximation defined as the hyperplane passing through the axis intercepts of the Gaussian constraint. Such a linear approximation is pictured in Fig. 6 and is equivalent to the *outer bound* approximation in [7].

By Theorem 1, the Gaussian acceptance region has a convex complement in the first quadrant, and therefore such a linear approximate acceptance region would contain points not contained in the Gaussian region. For this reason, we call this linear approximation the liberal linear constraint approximation. A conservative approximation would be a strict subset of the exact acceptance region.

An equation for this linear approximation is easily derived. Each axis intercept x_i^o of the Gaussian constraint is easily obtained by setting x_j $j \neq i$ to zero in (14) which yields

$$x_i^o = \frac{C}{\mu_i} - \frac{1}{2}\left(\frac{\rho\sigma_i}{\mu_i}\right)^2 \left(\sqrt{1+\frac{4C\mu_i}{\rho^2\sigma_i^2}} - 1\right)$$

$$= \frac{C}{\mu_i}\cdot\left[1-\frac{1}{2}K_i\left(\sqrt{1+\frac{4}{K_i}}-1\right)\right] \qquad (20)$$

where

$$K_i = \frac{\rho^2\sigma_i^2}{C\mu_i}.$$

The liberal linear constraint (LLC) and its normal vector are thus defined by

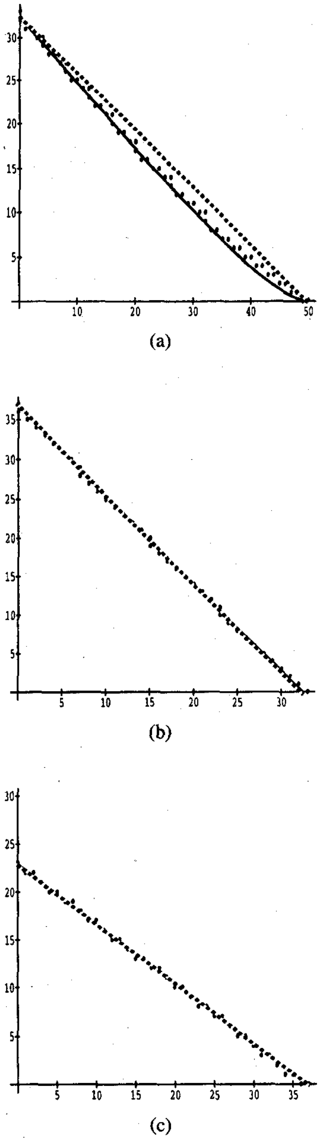
Liberal Linear Constraint Approximation:

$$LLC\colon \quad \sum_{i\in S} \frac{x_i}{x_i^o} - 1 = 0 \qquad (21)$$

$$\nabla\, LLC = \left(\frac{1}{x_1^o}, \frac{1}{x_2^o}, \cdots, \frac{1}{x_N^o}\right). \qquad (22)$$

The parameter K_i for each source is a function of the burstiness of the source and of the desired quality of service normalized by the capacity. These parameters will be useful in future sections. We note for sources with zero variance, $x_i^o = C/\mu_i$. As variance is increased, the intercepts decrease. We discuss sensitivities to system parameters in more detail in the next section.

In Fig. 7, we overlay the linear approximation (dashed line) with the previously generated Gaussian (solid curve) and exact (dots) constraints for the three examples of the previous section. The difference between the linear and Gaussian approximations is barely noticeable in Fig. 7(b) and (c), but significant in Fig. 7(a). See [7] for an analysis of the accuracy of this linear approximation.



(a)

(b)

(c)

Fig. 7. Liberal linear approximation to complete sharing acceptance region for three examples: (a) Type 1: deterministic, Type 2: on–off; (b) Type 1: on–off, Type 2: uniform; (c) Type 1: uniform, Type 2: arbitrary.

Note that the Gaussian boundary and linear constraint are equivalent if the two services are identical or if the means and variances of the services are scaled equally such that one service can be thought of as an additive collection of the other service. This implies if the burstiness descriptor σ_s^2/μ_s is the same for the two services, the Gaussian boundary will be linear.

Thus, if $\mu_1 = \alpha\mu_2$ and $\sigma_1^2 = \alpha\sigma_2^2$ for some $\alpha \in R_+$ then (15) becomes $C = (\alpha x_1 + x_2)\mu_2 + \rho\sqrt{\sigma_2^2(\alpha x_1 + x_2)}$ and the required capacity is constant along the linear boundary where $\alpha x_1 + x_2$ is itself a constant. This result holds for the general N source case as well.

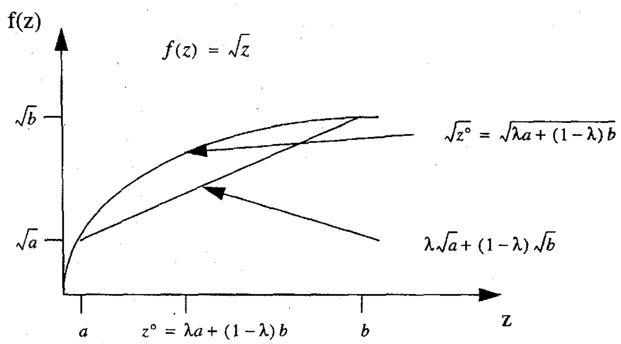

f(z)
$f(z) = \sqrt{z}$
$\sqrt{b}$
$\sqrt{a}$
$\sqrt{z^\circ} = \sqrt{\lambda a + (1-\lambda) b}$
$\lambda\sqrt{a} + (1-\lambda)\sqrt{b}$
a
$z^\circ = \lambda a + (1-\lambda) b$
b
z

Fig. 8. Graphic translation of constrained maximization problem of Theorem 2.

Now, hold the intercept point x_s° of the constraint fixed while increasing μ_s and decreasing σ_s^2 in proper proportion. The linear property of the constraint is destroyed and the Gaussian boundary diverges away from the linear constraint and bends in toward the origin. A point on the linear constraint will require a greater capacity than points on the Gaussian constraint. Denote the maximum of such additional capacity as $\Delta C_{\max}$, namely:

$$\Delta C_{\max} = \max\{C(X) - C \,|\, X \text{ satisfies (21)}\}.$$

$\Delta C_{\max}$ provides a measure of the distance between the Gaussian and liberal linear constraints as well as a measure of the cost incurred in multiplexing genuinely different source types. For this reason we call $\Delta C_{\max}$ the *diversity cost* of the system.

The diversity cost can be tightly bounded in the 2 service type case.

Theorem 2: In a $N = 2$ source type system, if x_1° and x_2° of type 1 and type 2 sources, respectively, require a capacity C at service criterion ρ then

$$\Delta C_{\max} = \frac{\rho(\sqrt{\sigma_2^2 \cdot x_2^\circ} - \sqrt{\sigma_1^2 \cdot x_1^\circ})^2}{4(\sqrt{\sigma_2^2 \cdot x_2^\circ} + \sqrt{\sigma_1^2 \cdot x_1^\circ})}.$$

Proof: From (15)

$$\begin{aligned} C &= \mu_1 x_1^\circ + \rho\sqrt{\sigma_1^2 \cdot x_1^\circ} \\ &= \mu_2 x_2^\circ + \rho\sqrt{\sigma_2^2 \cdot x_2^\circ}. \end{aligned} \tag{23}$$

Any point (x, y) on the liberal linear approximation can be represented as

$$(x, y) = \lambda(x_1^\circ, 0) + (1-\lambda)(0, x_2^\circ)$$

where $\lambda \in [0, 1]$. The additional capacity ΔC required at (x, y) is

$$\begin{aligned} \Delta C &= C(x, y) - C \\ &= C(\lambda x_1^\circ, (1-\lambda)x_2^\circ) - [\lambda C + (1-\lambda)C] \\ &= \rho\{\sqrt{\lambda\sigma_1^2 x_1^\circ + (1-\lambda)\sigma_2^2 x_2^\circ} \\ &\quad - [\lambda\sqrt{\sigma_1^2 x_1^\circ} + (1-\lambda)\sqrt{\sigma_2^2 x_2^\circ}]\}. \end{aligned}$$

Define $a = \min(\sigma_1^2 x_1^\circ, \sigma_2^2 x_2^\circ)$ and $b = \max(\sigma_1^2 x_1^\circ, \sigma_2^2 x_2^\circ)$.

$$\frac{\Delta C}{\rho} = \sqrt{\lambda a + (1-\lambda)b} - [\lambda\sqrt{a} + (1-\lambda)\sqrt{b}]. \tag{24}$$

$\Delta C/\rho$ is illustrated in Fig. 8 as the vertical distance between the curve and the line. This distance is maximum at a point $z^\circ = \lambda a + (1-\lambda)b$, where the derivative of $f(z)$ is equal to the slope of the line

$$\begin{aligned} \frac{1}{2}\frac{1}{\sqrt{z^\circ}} &= \frac{\sqrt{b} - \sqrt{a}}{b - a} \\ z^\circ &= \tfrac{1}{4}(\sqrt{b} + \sqrt{a})^2. \end{aligned}$$

The corresponding maximum vertical distance is

$$\frac{\Delta C_{\max}}{\rho} = \sqrt{z^\circ} - \left[\sqrt{a} + (z^\circ - a)\left(\frac{\sqrt{b} - \sqrt{a}}{b - a}\right)\right].$$

Some algebra yields

$$\begin{aligned} \Delta C_{\max} &= \frac{\rho}{4}\frac{(\sqrt{b} - \sqrt{a})^2}{\sqrt{b} + \sqrt{a}} \\ &= \frac{\rho(\sqrt{\sigma_2^2 \cdot x_2^\circ} - \sqrt{\sigma_1^2 \cdot x_1^\circ})^2}{4(\sqrt{\sigma_2^2 \cdot x_2^\circ} + \sqrt{\sigma_1^2 \cdot x_1^\circ})} \end{aligned}$$

or, alternatively

$$\Delta C_{\max} = \frac{(\mu_1 x_1^\circ - \mu_2 x_2^\circ)^2}{4\rho(\sqrt{\sigma_2^2 \cdot x_2^\circ} + \sqrt{\sigma_1^2 \cdot x_1^\circ})}.$$

In the 2 service type system, the diversity cost is an increasing function of the difference in burstiness of the two source types, or equivalently of the difference in mean aggregate source rates. If the two source types have identical burstiness ratios $\sigma_1^2/\mu_1 = \sigma_2^2/\mu_2$, then the cost of multiplexing the two types together, over that required if the entire capacity were devoted to either type, is zero. As the difference in burstiness increases this cost also increases.

The diversity cost has one additional use. A conservative linear constraint can be defined by first finding a capacity $C' < C$ such that $C' + \Delta C'_{\max} = C$ and then constructing the linear constraint in (21) with C replaced by C'. Such a constraint should be viewed as a non-unique linear constraint which is completely contained within the actual acceptance region and is consistent with the usual understanding of effective bandwidth for the unbuffered case [9]. In this view, the diversity cost may help us understand some of the potential inefficiencies of the effective bandwidth concept.

VI. Sensitivities

In this section, we use the Gaussian approximate resource allocation constraint to gain insight into the variation of the complete sharing acceptance region with source rate mean, variance, and quality of service. We present two methods of reacting to more demanding system parameters. First, we consider increasing capacity to accommodate the same number of each service type, then we consider decreasing the number of each service type under a constant capacity.

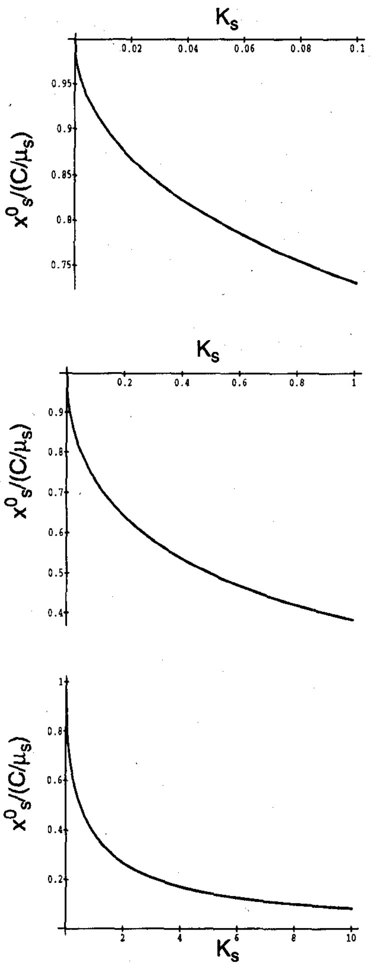


Fig. 9. $x_s^\circ/(C/\mu_s)$ plotted in terms of the parameter K_s.

First, consider the marginal capacity required to increase the rate mean or variance of source type s while maintaining a mix of services $X = (x_1, \cdots, x_N)$

$$\frac{\partial c}{\partial \mu_s} = x_s \tag{25}$$

$$\frac{\partial C}{\partial \sigma_s^2} = \frac{\rho x_s}{2\sigma_R(X)}. \tag{26}$$

The sensitivity of capacity to source rate mean is simply equal to the number of that source type in the service mix. The sensitivity to source rate variance is proportional to the number of that source type and to quality of service, and inversely proportional to the standard deviation of the total stream rate at the given operating point.

Next, consider the marginal capacity required to increase the quality of service while maintaining the same service mix

$$\frac{\partial C}{\partial PB_{\max}} = -\frac{\sigma_R(X)}{\varphi(\rho)}. \tag{27}$$

As the blocking probability decreases, ρ increases, and hence, $\varphi(\rho)$ decreases exponentially in ρ^2. As we would expect, this sensitivity is strongly dependent upon the Gaussian approximation, and might differ significantly if the tail of the distribution of the rate of the combined data stream has a different form.

Alternatively, we now consider keeping the capacity of the system constant, and reacting to more demanding system parameters by decreasing the acceptance region. The simplest manner of measuring such a decrease is to measure the change in the maximum number of each service type in the region, namely, the axis intercepts, x_s° defined in (20). The service mix could then be changed to any point remaining inside the corresponding acceptance region.

For a deterministic source, the axis intercept is equal to C/μ_s. As the burstiness of the source increases, the axis intercept decreases to some proportion of C/μ_s. Similarly, for any non-deterministic source, an increase in desired service quality produces a decrease of the axis intercept. The size of the decrease is entirely determined by the composite parameter K_s, as given by (20). Fig. 9 shows the relationship of $x_s^\circ/(C/\mu_s)$ with respect to the parameter $K_s = \rho^2\sigma_s^2/C\mu_s$ in three different ranges of K_s.

The decrease in the number of a service type that can be accommodated, in a system of capacity C with a quality of service given by ρ, when the source rate mean increases are given in (28)

$$\frac{\partial x_s^\circ}{\partial \mu_s} = -\frac{C}{\mu_s^2} - \frac{\rho^2\sigma_s^2}{\mu_s^3} + \frac{\dfrac{\rho^3\sigma_s^3}{\mu_s^4} + \dfrac{3C\rho\sigma_s}{\mu_s^3}}{\sqrt{\dfrac{\rho^2\sigma_s^2}{\mu_s^2} + \dfrac{4C}{\mu_s}}}. \tag{28}$$

The first term is the expected decrease due to the change in the ratio C/μ_s. The remaining, less significant terms reflect the secondary effect due to an increase in burstiness and hence, a change in K_s. This sensitivity is graphed in Fig. 10(a).

Similarly, the decrease in the number of a service type that can be accommodated in a system of capacity C with a quality of service given by ρ when the source rate variance increases is

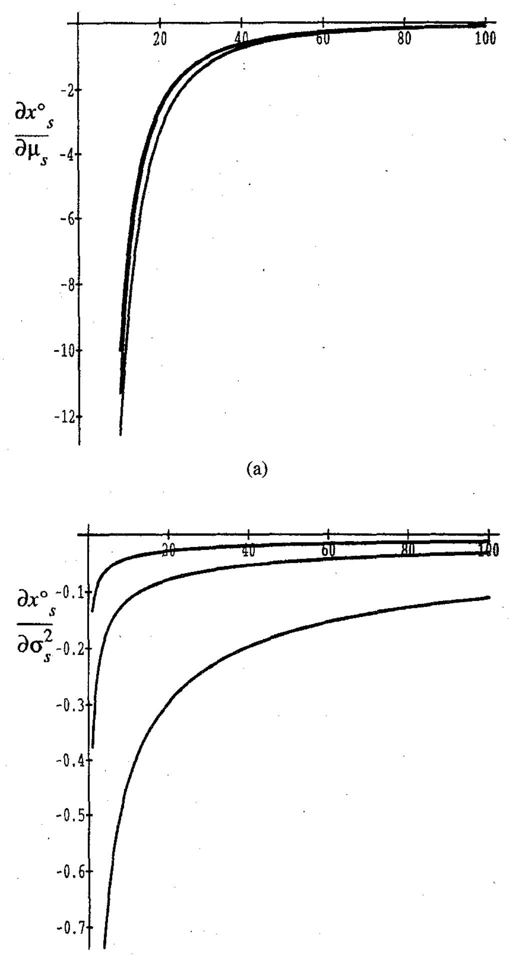
$$\begin{aligned}\frac{\partial x_s^\circ}{\partial K_s} &= -\frac{C}{2\mu_s}\left[\frac{K_s+2}{\sqrt{K_s^2+4K_s}} - 1\right]\\ \frac{\partial x_s^\circ}{\partial \sigma_s^2} &= \frac{\partial x_s^\circ}{\partial K_s}\cdot\frac{\rho^2}{C\mu_s}.\end{aligned} \tag{29}$$

Variance is increasing with K_s. For larger variances the quotient in the first line of (29) approaches one and the marginal number of service types approach zero asymptotically, This sensitivity is graphed in Fig. 10(b).

Finally, capacity and the maximum number of sources can be related by the marginal capacity required to increase the number of sources of type s, x_s, in a system operating at an operating point X and a blocking probability given by ρ

$$\frac{\partial C}{\partial x_s} = \mu_s + \frac{\rho\sigma_s^2}{2\sigma_R(X)}. \tag{30}$$

The first term is simply the extra capacity required to accommodate the extra mean bitrate. The second term is proportional to the desired quality of service and to the source type s rate variance normalized by the standard deviation of the total stream rate at the given operating point.



(a)

(b)

Fig. 10. Sensitivities of maximum number of sources to mean and variance. $C = 1000$, $PB_{\max} = 10^{-3}$. (a) $\sigma_s^2 = 0, 25, 50$ left to right; (b) $\mu_s = 100, 50, 10$ left to right. (a) Sensitivity to moon; (b) sensitivity to variance.

VII. Multiplexing Gains

In previous sections we have given approximate resource allocation constraints for multiplexing with complete sharing. In this section, we will consider similar approximate constraints for homogeneous sharing and no sharing multiplexing disciplines. Our primary concern is to analyze the gains achieved by various amounts of multiplexing.

We start with the Gaussian approximation to the complete sharing acceptance region. From (13), we recall that multiplexing of $X = (x_1, \cdots, x_N)$ sources under a complete sharing discipline at a quality of service given by ρ requires a capacity of approximately

$$C_{CS} \doteq \sum_{s=1}^{N} \mu_s x_s + \rho \sqrt{\sum_{s=1}^{N} \sigma_s^2 x_s}. \tag{31}$$

In homogeneous sharing, the channel is partitioned into sub-channels defined by source type. All sources of the type s are multiplexed onto sub-channel s, but no multiplexing is attempted between different source types. We assume the system accommodates a moderate number of each source type, and we use the same Gaussian approximation. Multiplexing of x_s sources of type s at a quality of service given by ρ thus requires a capacity of approximately

$$C_{HS,\,s} = x_s \mu_s + \rho \sqrt{\sigma_s^2 \cdot x_s} \quad \forall\, s \in S. \tag{32}$$

The total capacity required under a homogeneous sharing discipline to accommodate $X = (x_1, \cdots, x_N)$ sources at a quality of service given by ρ is, therefore, approximately

$$\begin{aligned} C_{HS} &= \sum_{s \in S} C_{HS,\,s} \\ &= \sum_{s=1}^{N} \mu_s x_s + \rho \sum_{s=1}^{N} \sqrt{\sigma_s^2 x_s}. \end{aligned} \tag{33}$$

In the no sharing discipline, there is no sharing of capacity between sources. Each source of type $s \in S$ is allocated a capacity of $C^\circ_{NS,\,s}$. For a single source, a Gaussian approximation would surely be erroneous, and therefore we simply assume that $C^\circ_{NS,\,s}$ is set such that each individual service of type s will satisfy the loss criterion $PB_{\max}$. Typically, we expect the capacity $C^\circ_{NS,\,s}$ to be near or equal to the peak rate of source type s when $PB_{\max}$ is small. The total capacity required under a no sharing discipline to accommodate $X = (x_1, \cdots, x_N)$ sources at a quality of service given by $PB_{\max}$ is therefore

$$C_{NS} = \sum_{s=1}^{N} C^\circ_{NS,\,s} x_s. \tag{34}$$

These three multiplexing disciplines are compared in Fig. 11 for the three example source mixes considered in previous sections. The acceptance regions, as calculated by (31), (32), (34), are shown in Fig. 11(b) and (c); (8), (34) are used in Fig. 11(a). In each case, the capacity allowed each discipline is fixed at $C_{CS} = C_{HS} = C_{NS} = 100$, and the quality of service is fixed at $PB_{\max} = 10^{-3}$. For the homogenous and no sharing disciplines, the boundary is mapped by varying the partition of the capacity among all possible combinations.

The complete sharing and homogeneous sharing discipline's boundary constraints intersect at axis intercepts, since the two schemes are identical at operating points where only one source type is accommodated. If there is, at most, one nondeterministic source type, e.g., Fig. 11(a), then the entire constraints overlap, since no meaningful sharing is possible. At operating points accommodating multiple source types, however, complete sharing results in capability to accommodate a greater number of each source type, providing there is some variability in at least one source rate. The distance between these two curves indicates the gain achieved by multiplexing heterogeneous source types. The Gaussian approximate acceptance region for complete sharing was shown to have a convex complement in the first quadrant in Theorem 1. Similarly, the Gaussian approximate acceptance region for homogeneous sharing has a similar character.

The no sharing boundary constraint is linear, from (34). The distance between this constraint and that corresponding to homogeneous sharing indicates the gain achieved by

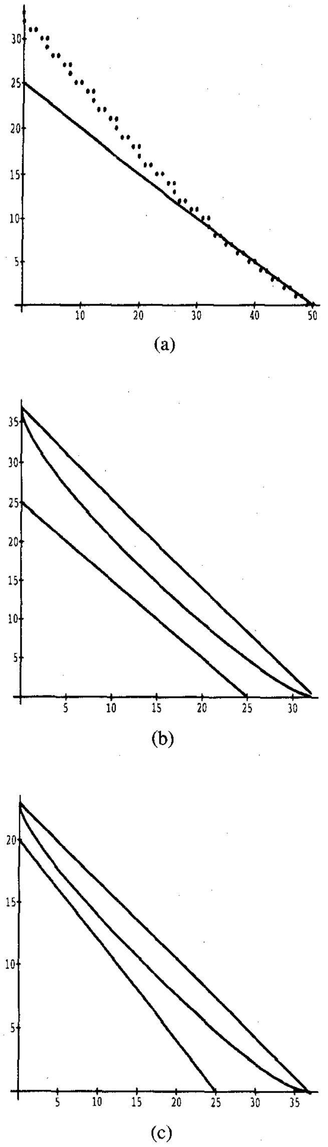


Fig. 11. Resource allocation constraints under (a) complete sharing; (b) homogeneous sharing and; (c) no sharing disciplines for three examples. Outside constraint is for complete sharing, middle is for homogeneous sharing, inside is for no sharing. (a) Type 1: deterministic, Type 2: on–off; (b) Type 1: on–off, Type 2: uniform; (c) Type 1: uniform, Type 2: arbitrary.

multiplexing homogeneous sources. This is the standard statistical multiplexing gain commonly addressed in the research literature.

To provide some insight into the variation of these two multiplexing gains with system parameters, we now compare competing multiplexing disciplines on the basis of the capacity required to accommodate a given combination X of service types at a given service quality $PB_{\max}$. This basis differs from that used in Fig. 11, where capacity was fixed and the acceptance regions were compared. This basis also differs from that used in the discussion of diversity cost in Section V. Diversity cost was defined for a single discipline, but for differing operating points. The comparisons in this section will be for a single operating point, but for differing disciplines.

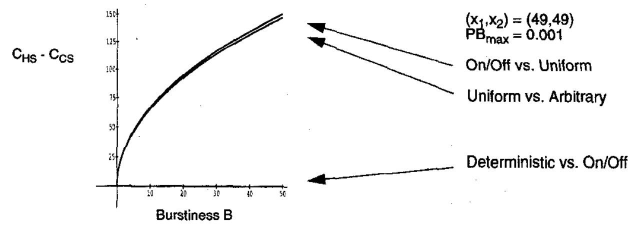


Fig. 12. Heterogeneous multiplexing gain for three examples.

An alternative basis might be the capacity required to accommodate a given combination X of service types at a given *loss probability*. The relationship between blocking probability, PB, and loss probability, PL, was given in (9) in general and for the Gaussian approximations. Cell loss probability is a more reasonable measure of quality of service than blocking probability, but results in a more complex comparison for the Gaussian case. Equating of blocking probabilities results in loss probabilities that are proportional to the coefficient of variation of the total rate of the multiplexed traffic. Since this coefficient of variation increases with partitioning of capacity, equating of loss probabilities would result in a greater required capacity than presented here for multiplexing disciplines that incorporate less sharing.

We know from the queuing literature that the magnitude of the statistical multiplexing gain is dependent on the variability of the customer process. We now consider the effect of burstiness on the homogeneous and heterogeneous multiplexing gains in this system. We define the burstiness of a source to be the source rate variance divided by the source rate mean

$$B_s = \frac{\sigma_s^2}{\mu_s}. \tag{35}$$

We consider increasing the burstiness of all source types simultaneously, in proportion, by defining a set of burstiness factors $p_1, p_2, \cdots, p_N$

$$B = p_1 B_1 = p_2 B_2 =, \cdots, = p_N B_N. \tag{36}$$

$C_{HS} - C_{CS}$ is the complete sharing gain achieved by sharing the entire channel between all sources over partitioning the channel among source types. From (31) and (32)

$$\begin{aligned} C_{HS} - C_{CS} &= \rho \left(\sum_{s=1}^{N} \sqrt{\sigma_s^2 x_s} - \sqrt{\sum_{s=1}^{N} \sigma_s^2 x_s} \right) \\ &= \rho\sqrt{B} \left(\sum_{s=1}^{N} \sqrt{\frac{\mu_s x_s}{p_s}} - \sqrt{\sum_{s=1}^{N} \frac{\mu_s x_s}{p_s}} \right). \end{aligned} \tag{37}$$

Note that $C_{HS} - C_{CS}$ is an increasing positive function of B, for any given service quality, operating point, mean source rates, and burstiness factors.

In Fig. 12, we illustrate the homogenous sharing gain for the three example illustrations at a specific operating point as we vary the burstiness of the sources using (37). Note that if one of the two source types is deterministic then p_s becomes arbitrarily large causing the multiplexing gain to be zero.

We define $C_{NS}-C_{HS}$ to be the multiplexing gain in sharing the channel among all services of the same type compared with allocating a fixed $C^{o}_{NS,\,s}$ for each type s service. Note that $C_{NS} - C_{HS}$ is a sum of multiplexing gains over each source type $s \in S$

$$C_{NS} - C_{HS} = \sum_{s=1}^{N} [C_{NS,\,s}x_s - (\mu_s x_s + \rho\sqrt{\sigma_s^2 x_s})]. \quad (38)$$

An upper bound on each term in the summation can be found by applying the Chebyshev Inequality

$$\begin{aligned} P(X-\mu > \varepsilon) &\le P(|X-\mu| > \varepsilon) < \frac{\sigma^2}{\varepsilon^2} \\ P(X > \varepsilon + \mu) &\le \frac{\sigma^2}{\varepsilon^2} \\ PB_{\max} &= P[r_{s,\,i}(t) > C^{o}_{NS,\,s}] \\ &\le \frac{B_s\mu_s}{(C^{o}_{NS,\,s} - \mu_s)^2} \\ C^{o}_{NS,\,s} &\le \left(\sqrt{B}\sqrt{\frac{\mu_s}{p_s PB_{\max}}}\right) + \mu_s. \end{aligned} \quad (39)$$

By substituting this result into (38), we find that the bound on $C_{NS}-C_{HS}$ also increases with the square root of burstiness B

$$C_{NS} - C_{HS} \le \sqrt{B}\sum_{s=1}^{N}\left[x_s\left(\sqrt{\frac{1}{PB_{\max}}} - \rho\sqrt{x_s}\right)\sqrt{\frac{\mu_s}{p_s}}\right].$$

In this section, we have seen that multiplexing is more beneficial when source types are more bursty. This does not imply that burstiness improves efficiency. Indeed, as source burstiness increases, all three curves shown in Fig. 3 move toward the origin (See [7] for results concerning burstiness versus channel utilization). The distance between each curve, however, also increases with burstiness. Therefore, if the system is required to transmit bursty sources, it is more efficient to share capacity as much as possible. Sharing within like source types results in a homogeneous sharing gain that increases with burstiness. This gain is what we traditionally call *statistical multiplexing gain.* Similarly, sharing between heterogeneous source types results in a complete sharing gain that also increases with burstiness. In both cases, the effect of multiplexing bursty sources is to reduce the burstiness of the combined stream and thereby increase the efficiency as given by channel utilization. Sources with less bitrate variance are already characterized by higher channel utilizations and the available multiplexing gain is therefore, diminished.

VIII. Summary

We have studied the variation of multiplexing gains with source burstiness in integrated service networks such as ATM. Using a Gaussian approximation for the distribution of bitrate of multiplexed sources, we considered three definitions of a virtual path in increasing order of bandwidth sharing: a single virtual circuit, all virtual circuits with identical paths and service types, and all virtual circuits with identical paths. We found that the capacity savings for systems that adopted greater sharing is proportional to the burstiness of the source types. These savings must be weighed against increased system complexity to accomplish greater sharing. Furthermore, we also found that the cost of having dissimilar source types present in the system is increasing with the difference between their burstinesses, suggesting that when bit streams are partially grouped, it is most important to first group similar source types.

The analysis is limited in several ways. First, the Gaussian approximation has been shown to be inaccurate in the tail, of interest when loss rates must be low. It is unknown whether the nature of the results found here will be affected by alternate tail behavior, such as that given by Chernoff bounds used by others to produce more accurate approximations to the acceptance region. Second, we assumed that all sources have an equal overflow probability. In ATM, virtual circuits with several different loss rate requirements might also be multiplexed onto a single virtual circuit. It would be of interest to similarly characterize the gains achieved by this additional level of resource sharing. Finally, the analysis is based on a system that uses only a small buffer, intended to resolve cell–scale congestion. It is unclear whether similar results would be found in a system that uses a large buffer, intended to resolve burst–scale congestion.

Appendix
Non-Gaussian Analysis

We consider in this appendix, the generalization of multiplexing gain results for non-Gaussian rate distributions. The capacity required under complete sharing, previously given in (31), now depends on the aggregate source rate distribution for all sources

$$C_{CS} = \sum_{s=1}^{N}\mu_s \cdot x_s + \rho(x_1, x_2, \cdots, x_N)\sqrt{\sum_{s=1}^{N}\sigma_s^2 \cdot x_s}$$

where

$$F_{R(X)}[\mu_{R(X)} + \rho(x_1, x_2, \cdots, x_N)\sigma_{R(X)}] = 1 - PB_{\max}.$$

Similarly, the capacity required under homogeneous sharing, previously given in (33), now depends on the aggregate source rate distribution for each source class

$$C_{HS} = \sum_{s=1}^{N}\mu_s \cdot x_s + \sum_{s=1}^{N}\rho_s(x_s)\sqrt{\sigma_s^2 \cdot x_s}$$

where

$$P\left(\sum_{j=1}^{x_s} r_{s,j}(t) > \mu_s + \rho_s(x_s)\sigma_s\right) = PB_{\max}\forall s \in S.$$

In the Gaussian case, $\rho(x_1, x_2, \cdots, x_N) = \rho_s(x_s) = \rho = \Phi^{-1}(1 - PB_{\max})$. Under general source rate distributions, this equality does not follow. Indeed, $\rho(x_1, x_2, \cdots, x_N)$ and $\rho_s(x_s)$ vary with the number of sources being multiplexed. As the number increases, the central limit theorem states that both quantities converge to $\rho = \Phi^{-1}(1 - PB_{\max})$.

The additional capacity required under homogeneous sharing over that required under complete sharing is thus given by

$$C_{HS} - C_{CS} = \sum_{s=1}^{N} \rho_s(x_s)\sqrt{\sigma_s^2 x_s} - \rho(x_1, x_2, \cdots, x_N)\sqrt{\sum_{s=1}^{N} \sigma_s^2 x_s}.$$

We consider an increase in the burstiness $B_s = \sigma_s^2/\mu_s$ of each source type. Now in the Gaussian case, $\rho(x_1, x_2, \cdots, x_N) = \rho_s(x_s) = \rho = \Phi^{-1}(1 - PB_{\max})$ is independent of burstiness. The resulting multiplexing gain thus follows from (37) either by fixing source means and varying source variances or vice versa.

In the general case, however, we must specify how to change the source rate distribution to effect an increase in the source burstiness. We classify such modifications on the basis of whether they effect $\rho(x_1, x_2, \cdots, x_N)$ and $\rho_s(x_s)$. Modifications that do not affect these parameters produce a multiplexing gain that varies with the square root of burstiness

$$C_{HS} - C_{CS} = \sqrt{B}\left(\sum_{s=1}^{N} \rho_s(x_s)\sqrt{\frac{\mu_s x_s}{p_s}} - \rho(x_1, x_2, \cdots, x_N)\sqrt{\sum_{s=1}^{N} \frac{\mu_s x_s}{p_s}}\right).$$

This class includes changes of scale and translations. Define H as a scaling of G: $F_H(t) = F_G(t/\alpha)$. Suppose $\rho_G = [F_G^{-1}(1 - PB) - \mu_G]/\sigma_G$, then

$$\rho_H = \frac{F_H^{-1}(1 - PB) - \mu_H}{\sigma_H} = \frac{\alpha(\mu_G + \rho_G\sigma_G) - \alpha\mu_G}{\alpha\sigma_G} = \rho_G$$

however, the burstiness has changed

$$B_H = \frac{\sigma_H^2}{\mu_H} = \frac{\alpha^2\sigma_G^2}{\alpha\mu_G} = \alpha B_G.$$

Similarly, define H as a translation of G: $F_H(t) = F_G(t - \beta)$, then

$$\rho_H = \frac{F_H^{-1}(1 - PB) - \mu_H}{\sigma_H} = \frac{\beta + (\mu_G + \rho_G\sigma_G) - (\beta + \mu_G)}{\sigma_G} = \rho_G.$$

Again, however, the burstiness has changed

$$B_H = \frac{\sigma_H^2}{\mu_H} = \frac{\sigma_G^2}{\beta + \mu_G}.$$

Therefore changes of scale and translations can be used to increase source burstiness, and such changes will produce multiplexing gains that are proportional to the square root of burstiness.

Modifications to the shape of a source rate distribution, however, modify $\rho(x_1, x_2, \cdots, x_N)$ and $\rho_s(x_s)$ and thus, produce multiplexing gains that may have a different dependence on burstiness.

References

[1] B. Bensaous, J. Guilbert, and J. W. Roberts, "Fluid queueing models for a superposition on on–off sources," presented at ITC Seminar, Morristown, NJ, 1990.

[2] M. Butto, E. Cavellaro, and A. Tonietti, "Effectiveness of the leaky bucket policing mechanism in ATM networks," *IEEE J. Select. Areas Commun.* vol. 9, pp. 335–342, 1991.

[3] A. I. Elwalid and D. Mitra, "Analysis and design of rate-based congestion control of high speed networks, I: Stochastic fluid models, access regulations," *Queueing Systems*, vol. 9, pp. 29–64, 1991.

[4] ——, "Effective bandwidths of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 329–343, 1993.

[5] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Syst.*, vol. 9, pp. 17–28, 1991.

[6] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high speed network," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 968–981, 1991.

[7] J. Y. Hui, "Resource allocation for broadband networks," *IEEE J. Select. Areas Commun.* vol. SAC-6, pp. 1598–1608, 1988.

[8] ——, *Switching and Traffic Theory for Integrated Broadband Networks.* Norwell, MA: Kluwer, 1990.

[9] F. P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Syst.*, vol. 9, pp. 5–15, 1991.

[10] G. Kesidis and J. Walrand, *Traffic Policing and Enforcement of Effective Bandwidth Constraints in ATM Networks*, preprint.

[11] 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, Aug. 1993.

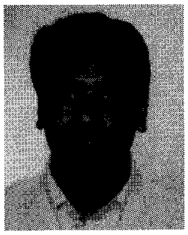
[12] D. Mitra and J. Morrison, "Erlang capacity and uniform approximations for shared unbuffered resources," submitted to *IEEE/ACM Trans. Networking*.

[13] I. Norros *et al.*, "The superposition of VBR sources in an ATM multiplexer," *IEEE J. Select. Areas Commun.*, vol. 9, no. 3, pp. 378–387, 1991.

[14] P. Rathgeb, "Modeling and performance comparisons of policing mechanisms for ATM networks," presented at ITC Seminar, Morristown, NJ, 1990.

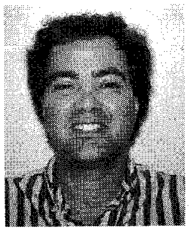
[15] J. Roberts, "Variable bitrate congestion control in B-ISDN," *IEEE Commun. Mag.*, vol. 29, no. 9, pp. 50–56, 1991.

[16] K. Sohraby and M. Sidi, "On the performance of bursty and correlated sources subject to leaky bucket rate-based access control schemes," in *Proc. IEEE Infocom*, (4D.3), Apr. 1991, pp. 426–434.



Ikhlaq Sidhu (S'93–M'95) received the B.S.E.E. degree from the University of Illinois, Urbana–Champaign, and the M.S.E.E. and Ph.D. degrees from Northwestern University, Evanston, IL, in 1995.

He is a researcher and member of the architecture group of the Corporate Systems Division for US Robotics in Skokie, IL. His research interests include traffic modeling, call admission control, and architectures for network access.



Scott Jordan (S'83–M'90) received the B.S., M.S., and Ph.D. degrees from the University of California, Berkeley, in 1985, 1987, and 1990, respectively.

He is currently an Assistant Professor at Northwestern University, Evanston, IL. His teaching and research interests are the modeling and analysis of behavior, control, and pricing of computer/telecommunication networks.