Extending the Effective Bandwidth Concept to Networks with Priority Classes

Arthur W. Berger, Lucent Technologies Bell Laboratories
Ward Whitt, AT&T Laboratories

ABSTRACT ATM switches are now being designed to allow connections to be partitioned into priority classes, with packets being emitted for higher priority classes before packets are emitted for lower priority classes. Accordingly, allocation of network resources based on different priority levels is becoming a realistic possibility. Thus, we need new methods to do connection admission control and capacity planning that take account of the priority structure. In this article we show that the notion of effective bandwidths can be used for these purposes when appropriately extended. The key is to have admissibility of a set of connections determined by a linear constraint for each priority level, involving a performance criterion for each priority level. For this purpose, connections are assigned more than one effective bandwidth, one for its own priority level and one for each lower priority level. Candidate effective bandwidths for each priority level can be determined by using previous methods associated with the first-in first-out discipline. The proposed effective bandwidth structure makes it possible to apply product-form stochastic loss network models to do dimensioning.

Emerging high-speed communication networks, such as broadband ISDN networks that employ ATM technology, tend to be packet networks rather than circuit-switched networks because the packet structure allows for better resource sharing. In a packet network, sources do not require dedicated bandwidth (e.g., circuits) for the entire duration of a connection. Unfortunately, however, the enhanced flexibility of packet networks also makes it more difficult to effectively control the admission of connections seeking to enter an existing network, and to plan the capacity of future networks when they are designed.

The problems of admission control and capacity planning in a packet network may be addressed by a concept known as the *effective* or *equivalent bandwidth* of a connection. When employing this concept, an appropriate effective bandwidth is assigned to each connection, and each connection is treated as if it required this effective bandwidth throughout the active period of the connection. The feasibility of admitting a given set of connections may then be determined by ensuring that the sum of the effective bandwidths is less than or equal to the total available bandwidth (i.e., the capacity). By using effective bandwidths in this manner, the problems of admission control and capacity planning are addressed in a fashion similar to that employed in circuit-switched networks.

Of course, the actual bandwidth (bit rate) needed by each variable bit rate (VBR) connection is uncertain and fluctuates over time, as depicted in Fig. 1. The actual required bandwidth fluctuates between some minimal level, perhaps 0, and a peak rate, which is typically determined by the speed of the access line. It is evident that the effective bandwidth of a connection should be some value between its average rate and its peak rate. Any particular value that is used is necessarily an approximation, but potentially a very useful approximation. Let e_i be the effective bandwidth assigned to a connection of type i; let n_i be the number of connections of type i; let I be the number of connection types; and let c be the capacity of a link; for ATM networks the link could be replaced by a virtual path (VP). The set of connections determined by the vector $(n_1, ..., n_I)$ is said to be *admissible* if

$$\sum_{i=1}^{I} e_i n_i \le c. \qquad (1)$$

When the network contains multiple constrained resources, there is such a constraint for each resource. Then a set of connections is deemed admissible if inequality Eq. 1 holds for each resource. A candidate new connection is admitted if the set of existing connections plus the new connection produces a feasible set of connections. Otherwise, the candidate new connection is rejected.

To do capacity planning or dimensioning, we can specify arrival rates and average holding times for each connection type. Then, assuming a product-form stochastic loss network model as in Ross [1], we can compute blocking probabilities for each connection type for any given capacity. These blocking probability calculations can be efficiently performed by numerically inverting the generating function of the normalization constant in these product-form models, as in Choudhury, Leung, and Whitt [2]. We then choose the capacity of each resource so that the blocking probabilities are suitably small. Moreover, we need not use a complete-sharing policy. We can improve performance by imposing upper-limit and guaranteed-minimum constraints on the connection classes. An upper limit of U_i on type i restricts the number of type-i connections that can be simultaneously present to be at most U_i. A guaranteed minimum of M_i for type i is a constraint on all other types, ensuring that there is always room for at least

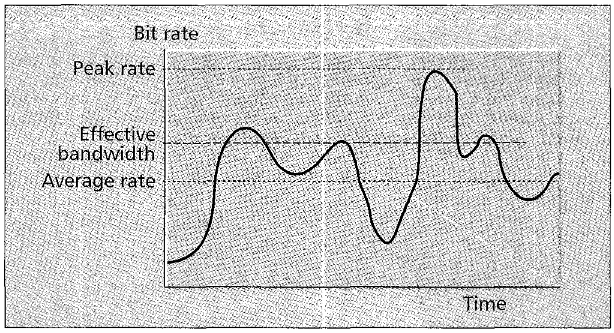
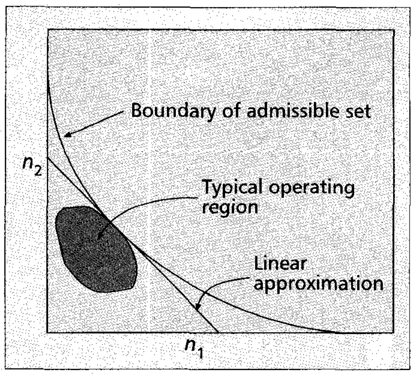


Figure 1. *Instantaneous rate of a VBR connection.*

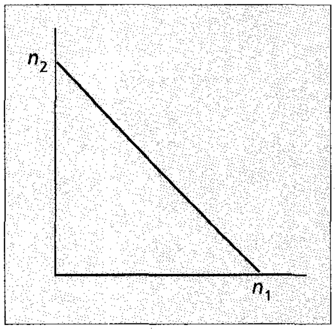
0163-6804/98/$10.00 © 1998 IEEE

M_i type-i connections. With these constraints, the blocking probabilities can still be efficiently computed by numerical inversion. Moreover, the capacities and sharing parameters can be found by a search algorithm, as in [3].



■ **Figure 2.** *A nonlinear admissible set for two classes, and an approximating linear admissible set.*

Over the last ten years considerable work has been done on effective bandwidths. A theoretical basis for Eq. 1 was developed in the context of large-buffer large-deviation asymptotics. Recent reviews can be found in Chang and Thomas [4], de Veciana, Kesidis, and Walrand [5], and Kelly [6]. Unfortunately, however, the effective-bandwidth approach based completely on large-buffer asymptotics is often not a very accurate approximation; see Choudhury, Lucantoni, and Whitt [7]. Hence, various refinements have been proposed, many abandoning the linear structure Eq. 1. However, as indicated above, the linear structure in Eq. 1 can greatly assist engineering. Thus, we propose keeping Eq. 1 and allowing the effective bandwidths e_i to be adjusted as needed. In particular, given a nonlinear admissible set associated with some other admission control procedure, we can obtain effective bandwidths by introducing a linear approximation to the nonlinear admissible set. To do so, we might exploit knowledge of the typical operating region. For example, consider the case of two classes. A nonlinear admissible set might look as depicted in Fig. 2. We might know that the typical operating region is the shaded region in Fig. 2. Then we might approximate the admissible set by a linear hyperplane, chosen to be tangent to the admissible set at a point near the typical operating region. This line implicitly defines effective bandwidths for the two classes. In particular, the approximate effective bandwidths are $e_i^* = c/n_i^*$ where n_i^* is the point on the n_i axis intersected by the tangent line.



■ **Figure 3.** *An admissible set with two classes.*

The concept of effective bandwidths has been developed for buffers using the first-in first-out (FIFO) service discipline. However, now ATM switches are being designed to allow the connections to be partitioned into priority classes with packets being emitted from higher priority classes before lower priority classes. This priority structure is useful to meet the different requirements of the diverse traffic that will be carried at ATM networks. Typical implementations have from two to four priority classes. The highest priority class might be constant bit rate (CBR) traffic. The next priority class might be real-time (interactive) video traffic. Non-real-time VBR traffic could be a lower priority class, which might be further divided into two priorities, making a lowest priority class for best-effort or available bit rate (ABR) traffic.

It is natural, then, to consider how the concept of effective bandwidths should be modified to properly take account of priority classes; that is the topic of this article.

Modifications of Effective Bandwidths for Priorities

In this section we present an informal engineering argument to show that, regardless of the method used for computing effective bandwidths, if priorities are implemented in the network node, then practical and efficient engineering rules should use a linear constraint per-priority (leading to a trapezoidal admissible set for two priorities) and wherein a given connection type is associated with *multiple* effective bandwidths.

Before doing so, we point out that a more formal mathematical development based on large-buffer asymptotics is presented in [8]. There we show that the admissible set resulting from full asymptotic analysis does *not* actually have the proposed structure, but that a reasonable approximation does. We also review the related literature in [8]. Here we simply point out that other researchers previously began to examine the impact of non-FIFO queuing on bandwidth allocation and admission control in high-speed networks. See de Veciana and Kesidis [9] and Zhang [10] for the generalized processor-sharing policy (which contains the two-priority model as a special case) and Elwalid and Mitra [11] and Kulkarni and Gautam [12] for priority disciplines. We and the others consider the case in which each class has its own queue with its own buffer. The analyses in [8–12] can be interpreted as providing additional support for our proposal.

To consider how effective bandwidths might be extended to accommodate priority classes, consider the simple case of a single link with two connection types, with type 1 having priority over type 2. Before introducing priorities, the admissible set is the set of pairs (n_1, n_2) such that $e_1n_1 + e_2n_2 \le c$, where e_i is the effective bandwidth of class i and c is the capacity (output rate); that is, with the FIFO discipline we have a triangular admissible set as depicted in Fig. 3.

The primary reason we should want to use priority service is that the lower priority class has a looser performance criterion than the higher priority class. To be more precise, suppose that each priority class has its own buffer, with the class i buffer having capacity b_i. The performance criterion for class i might be that the long-run proportion of cells lost due to buffer overflow be less than p_i. The class 2 criterion might be weaker because b_2 is greater than b_1 or because p_2 is greater than p_1, or both.

The looser criterion for class 2 means that the effective bandwidth for a class-2 connection should be less with the priority discipline than with FIFO. If we considered class 2 alone, then we should be able to admit more class 2 connections when we introduce the looser constraint. However, the more stringent class 1 constraint would leave the admissible set unchanged away from the class 2 axis. However, when we introduce priorities along with the looser class 2 constraint, we expect that the FIFO admissible set depicted by triangle A in Fig. 4 should be replaced by the larger triangle $A + B$ in Fig. 4.

However, introducing priorities has a further impact. Upon further reflection, we realize that there should still be some bandwidth left over for class 2 when the high-priority class 1 has "filled the link." Of course, if the high-priority class is CBR traffic or nearly CBR traffic, then its peak rate would be very close to its average rate, so that there would be negligible room for class 2 when class 1 reaches its performance limit.

However, if class 1 traffic has considerable variability, then its peak rate might be much greater than its average rate, so that there might be considerable room in the link in terms of average rate when the class 1 priority limit is reached. For example, in ATM if the VBR real-time connections have filled the link according to their effective bandwidth, there would likely be room for some lower-priority ABR connections. Indeed, the occupancy might well be 50 percent or less when the VBR traffic is at its upper limit. Even with CBR high-priority connections, there may be some spare bandwidths for lower-priority connections, because performance criteria on cell delay variation might limit the occupancy of CBR connections to, say, 90 percent.

Given that some class 2 connections can be admitted when class 1 is at its upper limit, we expect a vertical segment on the right of the admissible set. Instead of the triangular admissible $A + B$ set in Fig. 4, we should anticipate the trapezoidal admissible set $A + B + C$ in Fig. 4. In order to have the trapezoidal admissible set in Fig. 4, we need a second linear constraint. We now should have the pair of constraints

$$\begin{aligned} e_1 n_1 &\leq c \\ e_1^2 n_1 + e_2 n_2 &\leq c. \end{aligned} \qquad (2)$$

The first constraint is the same constraint for class 1 with FIFO. The second constraint is the new linear constraint, which incorporates the reduced effective bandwidth e_2 for class 2 due to its looser performance constraint. The new parameter e_1^2 is determined by the height of the vertical segment on the right of the trapezoidal admissible set in Fig. 4.

In constructing the trapezoidal admissible set in Fig. 4, we have assumed that we know the class-2 limits when class-1 is at its lower and upper limits. The linearity in between can be regarded as the effective-bandwidth approximation.

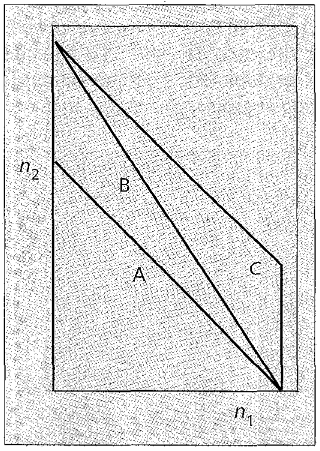
It is useful to interpret the new parameter e_1^2 in Eq. 2. The parameter e_1^2 can be regarded as the effective bandwidth for a priority-1 connection that is subject to the priority-2 performance criterion. We say that e_1^2 is the effective bandwidth for a priority-1 connection *as seen by* priority 2. Given the sensible case in which the priority 1 criterion is tighter than the priority 2 criterion, we have

$$e_1^2 < e_1. \qquad (3)$$

In the construction of Fig. 4 from Fig. 3, we relied on the inequality Eq. 3. If instead we have $e_1 < e_1^2$, then the first constraint in Eq. 2 would be vacuous.

Figure 4 is useful to graphically see the advantage of introducing priority classes. The change from triangle A to trapezoid $A + B + C$ shows the gain achieved from introducing priorities. Constructing the admissible sets with and without priority classes can be very helpful to see the advantage of having priorities, where in the case of no priorities (FIFO service) the admission of connections of any type would be subject to the strictest (otherwise priority 1) performance criterion. In some cases priorities may provide a big gain, while in other cases they may only provide a modest gain.

The notion of per-priority effective bandwidth generalizes to an arbitrary number of priority classes. For three priority classes, the admissible set is



■ **Figure 4.** *An admissible set with two priority classes. Class 1 has priority over class 2.*

$$\begin{aligned} e_1^1 n_1 &\leq c \\ e_1^2 n_1 + e_2^2 n_2 &\leq c \\ e_1^3 n_1 + e_1^3 n_2 + e_3^3 n_3 &\leq c, \end{aligned} \qquad (4)$$

where e_i^k is the effective bandwidth for a priority-class-i connection as seen by priority k, with $i \leq k$ in all cases. In Eq. 4 we have used e_i^i to denote e_i. Multiple connection types within a given priority class are treated just as with FIFO. Let i denote the priority level and let j denote the connection type, where $1 \leq j \leq J_i$ and $1 \leq i \leq I$. Let e_{ij}^k denote the effective bandwidth of a priority-i type-j connection as seen by priority k. We need e_{ij}^k only for $k \geq$ i. With I priority levels, the admissible set is determined by the I constraints

$$\sum_{i=1}^{k} \sum_{j=1}^{J_i} e_{ij}^k n_{ij} \leq c, \qquad k = 1,\ldots,I. \qquad (5)$$

The sum over i in Eq. 5 could be extended to all i (up to I) provided that we set $e_{ij}^k = 0$ for $k < i$.

Loss Versus Delay Performance Criteria

There are two different performance criteria that are commonly considered: cell loss probabilities and delay tail probabilities. With the FIFO discipline, these two criteria are closely related. It is common to use the tail probability of the queue-length distribution in an unlimited-buffer model to approximate the cell loss probability. However, assuming constant-size cells, the delay at any time is a constant multiple of the queue length. Thus, with the FIFO discipline and an unlimited-buffer approximation, any delay performance criterion translates into an equivalent cell loss probability requirement.

However, with priority classes the equivalence between delay and cell loss no longer holds. A lower priority class cell has to wait not only for all cells of its priority and all higher priorities that are currently in the system; it also has to wait for new higher-priority cells that arrive *after* the lower-priority cell arrives, but before it can receive service. Thus, the delay can be much greater than determined by the workload vector seen upon arrival.

Thus, to be specific and to avoid confusion, previously we stipulated that each priority class had its own buffer and that the performance criterion for each class was based on the cell loss probability. However, our approach to effective bandwidth with priorities is quite general, so that it should accommodate variations in the model and performance criteria.

Determining the Effective Bandwidths

So far, our analysis holds independently of how the given effective bandwidths are calculated. In the context of FIFO service, various methods have been proposed for making such calculations; see Kelly [6] for a nice overview. Any of these methods can be extended to incorporate the per-priority effective bandwidths proposed herein. The idea is to produce a linear constraint for each priority class, based on the performance criterion for that class, and considering only connections from that class and all higher priority classes. Any method for calculating the admissible set with the FIFO discipline can be used for each constraint, provided we introduce a linear approximation whenever it is needed.

In the present section we present one method that can be used in the FIFO context and show how it can be adapted to priority service. Three other methods are described in [8]: one is based on large-buffer asymptotics; a second is based on measuring the admissible set considering two connection types at a time; and a third is based on a standardized traffic descriptor. All these methods, both in the FIFO context and in the generalization to priorities, exploit the assumed linearity in the effective bandwidth constraints, Eqs. 1 and 5.

The complexity (e.g., long-range dependence) revealed by many network traffic measurements (e.g., see [13] and references therein) suggest that it may be desirable to use a conservative bufferless model for connection admission control and/or capacity planning. We describe one way that effective bandwidths can be defined for such a model in this subsection.

We first consider the FIFO case and suppose that source i sends cells, modeled as fluid, at a random rate $R_i(t)$, which has mean m_i and variance σ_i^2. Thinking of the case in which there are many sources, each small compared to the total, we invoke the central limit theorem (CLT) to justify approximating the aggregate rate $R(t) = R_1(t) + \ldots + R_n(t)$ from n sources as a Gaussian process with mean $m = \Sigma_{i=1}^{n} m_i$ and variance $\sigma^2 = \Sigma_{i=1}^{n} \sigma_i^2$.

We consider a collection of sources admissible if the probability that the total input rate exceeds the constant output rate c (or some other target level) is suitably small, i.e., if

$$P(R(t) > c) \le p \tag{6}$$

for appropriate p. Using the Gaussian approximation, the admissible set for J types of sources is the set of vectors

$$\left\{(n_1,\ldots,n_J): \sum_{j=1}^{J} n_j m_j + \alpha\sqrt{\sum_{j=1}^{J} n_j \sigma_j^2} \le c\right\}, \tag{7}$$

where n_j is the number of sources of type j and $\Phi^c(\alpha) \equiv P(N(0,1) > \alpha) = p$ with $N(0,1)$ being a standard (mean 0, variance 1) normal random variable. If very small probabilities p are considered in Eq. 1, then it may be better to replace the Gaussian CLT approximation by a large-deviations approximation. Alternatively, we might make the CLT approximation more suitable by choosing a target rate c in Eq. 6 below the output rate, which allows us to increase the target probability p.

It should be noted that the constraint in Eq. 7 is quadratic in the variables $\sqrt{n_j}$, so that the admissible set does not directly have linear boundaries. Nevertheless, we can produce an approximate admissible set with a single linear boundary. One way to do so is to consider the classes one at a time. For source type j, we solve the quadratic equation

$$n_j m_j + \alpha\sqrt{n_j \sigma_j^2} = c \tag{8}$$

to obtain the maximum number $\bar{n}_j$ of type j connections when only type j is present, rounding down to produce an integer. We then let the effective bandwidth be defined by

$$e_j = \frac{c}{\bar{n}_j}. \tag{9}$$

Unfortunately, however, the admissible set Eq. 7 is a proper subset of the admissible set based on Eq. 9. Thus, to be more conservative, we might want to increase the effective bandwidths defined by Eq. 9. A simple way to do this is to modify the effective bandwidths in Eq. 9. If we think type j will constitute a proportion p_j of the total number of connections, we can solve the single quadratic equation

$$n\left(\sum_{j=1}^{J} p_j m_j\right) + \alpha\sqrt{n\sum_{j=1}^{J} p_j \sigma_j^2} = c \tag{10}$$

for n^*, the total number of feasible connections, assuming the distribution $(p_1, \ldots, p_J)$ among the J types. We then can increase the effective bandwidths in Eq. 9 for each type j by setting

$$e'_j = \frac{c e_j}{\sum_{l=1}^{J} n^* p_l e_l} \tag{11}$$

where e_j is given in Eq. 9.

A more conservative procedure is to use the tangent hyperplane to the admissible set Eq. 7 at the point $\hat{n} \equiv (n^* p_1, \ldots, n^* p_J)$ where n^* is the solution to Eq. 10 and $(p_1, \ldots, p_J)$ is the assumed vector of proportions. The tangent hyperplane to the admissible set Eq. 7 at $\hat{n}$ is determined by the equation

$$\nabla f(n)|_{\hat{n}} \cdot (n - \hat{n}) = 0, \tag{12}$$

where ∇ is the gradient and f is the function on the left of the inequality in Eq. 7, regarding n and $\hat{n}$ as vectors of continuous variables. Since

$$\left.\frac{\partial f}{\partial n_j}\right|_{\hat{n}} = m_j + \frac{1}{2}\alpha\sigma^2\left(\sum_{l} \hat{n}_l \sigma_l^2\right)^{-1/2}, \tag{13}$$

the effective bandwidth of a type-j source in the tangent-hyperplane admissible set is

$$e_j^* = \frac{(m_j + \sigma_j^2 \delta)c}{\sum_{l=1}^{J} (m_l + \sigma_l^2 \delta)\hat{n}_l}, \tag{14}$$

where

$$\delta = \frac{1}{2}\alpha\left(\sum_{j=1}^{J} \hat{n}_j \sigma_j^2\right)^{-1/2} \tag{15}$$

The bufferless model also helps confirm our main conclusions about effective bandwidths with priorities. If we use the bufferless model with I priority classes, then instead of Eq. 6 we directly obtain the I constraints

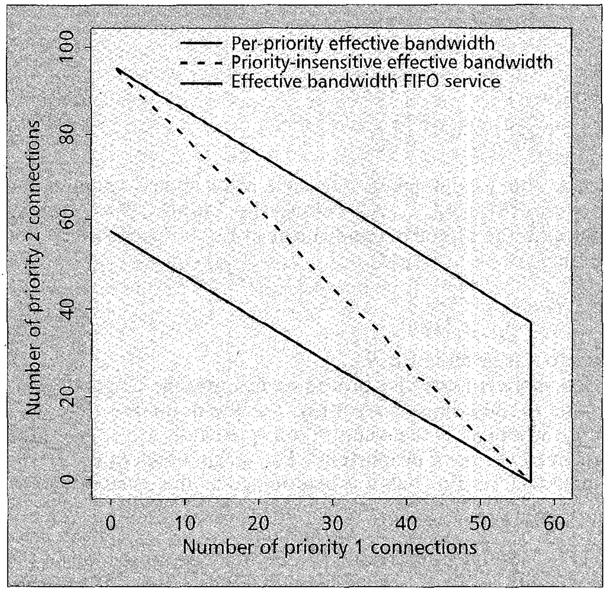
$$P(R^{(1)}(t) + \ldots + R^{(k)}(t) > c) \le p_k, \quad 1 \le k \le Z, \tag{16}$$

where $R^{(i)}(t)$ is the total input rate for priority class i. We obtain the multiple-effective-bandwidth structure proposed earlier if we approximate each of the I constraints in Eq. 16 by a linear constraint, which might be done in the manner described above.

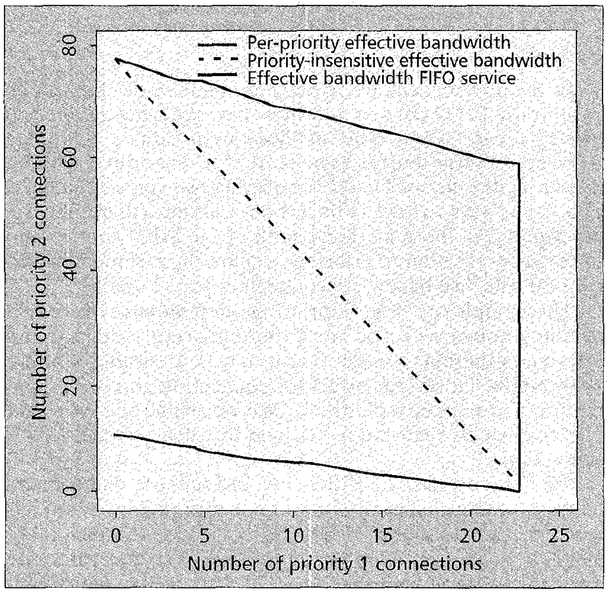
Numerical Examples

In this section we present four numerical examples to illustrate the benefits from using the per-priority effective bandwidths, including a case where the benefits are only marginal. A different issue is the accuracy of the effective-bandwidth approximations, of whatever type, as compared with the "exact" calculation of the admissible set. We do not pursue that issue here, but a detailed discussion is given in [8] for the large-buffer-asymptotic method.

For the examples, we use effective bandwidths based on large-buffer asymptotics assuming a fluid queue model with an infinite buffer, as discussed in [8]. For simplicity, we consider two priorities and one type of connection in each priority. For the first example we start with the simple case in which the connections are the same for each priority. (This could represent the case in which some users are given better ser-



■ **Figure 5.** *Example 1.*

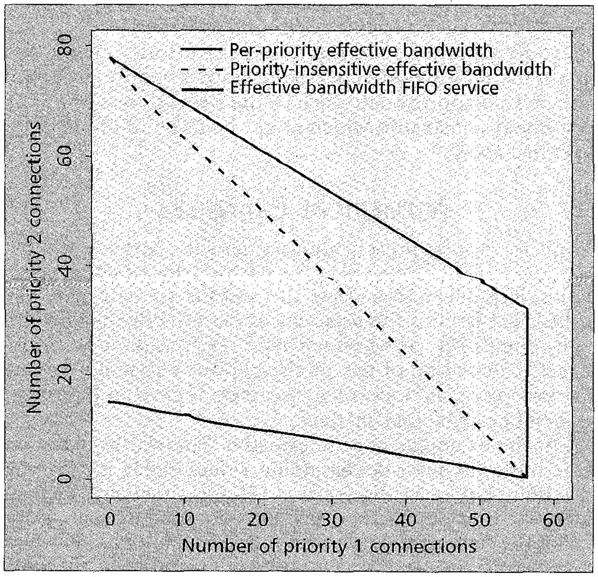


■ **Figure 7.** *Example 3.*

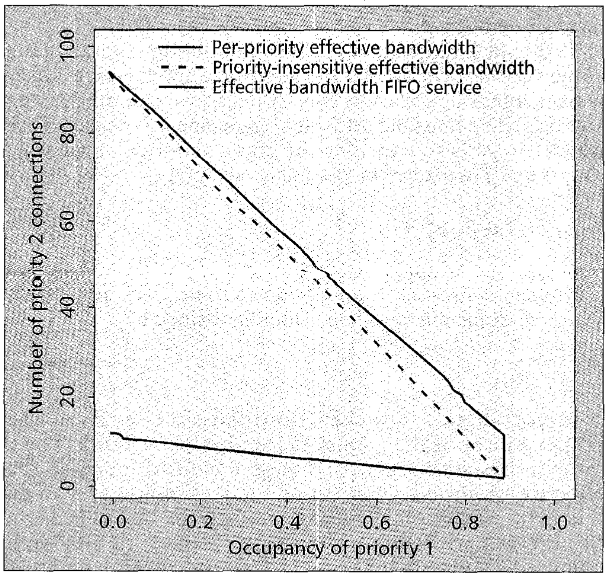
vice for a higher price.) Suppose that the connections are modeled as on-off two-state Markov modulated Poisson processes (MMPPs), where each arrival offers one unit of work (corresponding to one ATM cell). Suppose that the mean rate is 0.01, the fraction of time on is 0.1, and the mean burst size is 20. Let the performance criterion for priority class i require that the probability that the steady-state buffer content should exceed a level b_i be less than p_i. In this example, let the performance parameters be: $b_1 = 500$, $b_2 = 5{,}000$, and $p_1 = p_2 = 10^{-6}$. Lastly, let the link bandwidth be 1, which is 100 times the mean rate. For these parameters, the effective bandwidths turn out to be $e_1 = 0.0174$, and $e_1^2 = e_2 = 0.0105$ (see [8]). Note that since the connection type is the same for both priorities, e_1^2 equals e_2. Also note that e_1 is larger than e_1^2, and that the priority 2 performance criterion parameters are qualitatively looser than priority 1's.

The admissible set for three cases is given in Fig. 5. The smallest admissible set (lower solid line) labeled "Effective bandwidth FIFO service" assumes that priority service has *not* been implemented, the service discipline is FIFO, and the stricter performance criterion applies to all connections. In this case, e_2 would equal e_1, which is 0.0174, and the admissible set is given by Eq. 1 with $I = 2$. The middle admissible set (dashed line), labeled "priority-insensitive effective bandwidth," assumes priority service *has* been implemented, and the looser performance criterion applies to priority 2, but just one effective bandwidth, e_1, is used for the priority 1 connections. Again the admissible set is given by Eq. 1 with $I = 2$. The largest admissible set (upper solid line), labeled "per-priority effective bandwidth," uses two effective bandwidths, e_1 and e_1^2, for the priority 1 connections, and the admissible set is given by Eq. 2.

The main point of Example 1 is that we see the phenomenon depicted in Fig. 4. Note especially that at higher occupancies of priority 1, the admission of priority 2 connec-



■ **Figure 6.** *Example 2.*



■ **Figure 8.** *Example 4.*

tions is needlessly limited if the effective bandwidths are not adjusted for priorities. For example, when n_1 is 50, the priority-insensitive effective bandwidths limit n_2 to 12, whereas n_2 would be 45 with per-priority effective bandwidths. If no priorities are used, n_2 is 7. Note that half of the potential gain (measured in terms of area of admissible sets) from implementing priorities is not realized if the effective bandwidths are priority-insensitive.

Example 2 is the same as Example 1, except that the priority 2 connections are more bursty: the mean burst size is changed from 20 to 100. Then e_2 changes from 0.0105 to 0.0128. The resulting admissible sets are shown in Fig. 6. As in Example 1, at higher occupancies of priority 1, the admission of priority 2 connections is needlessly limited if the effective bandwidths are not adjusted for priorities. Also, in the present example we see more gain from the implementation of priorities than in Example 1. For instance, in Example 2 when n_1 is small, say zero, the looser criterion used for priority 2 allows 5.2 times more connections to be admitted as compared with FIFO service. In Example 1, this factor was "only" 1.7.

The occupancy on the link due to the priority 1 connections influences the potential gain from the per-priority effective bandwidths. In Examples 1 and 2, when the number of priority 1 connections admitted is the maximum possible and no priority 2 connections are present, the occupancy is 57 percent. Example 3 considers the case where this maximum priority 1 occupancy is lower, and Example 4 considers the case where it is higher.

Example 3 is the same as 2 except the priority 1 performance criterion is tighter: b_1 is reduced from 500 to 200, which is still ten times greater than the mean burst size. The resulting admissible sets are given in Fig. 7. In Example 3 the maximum number of priority 1 connections admissible is 23, and thus the maximum priority 1 occupancy is only 23 percent. Here we see a very strong advantage of using per-priority effective bandwidths. For instance, when n_1 is 20, the priority-insensitive effective bandwidths restrict n_2 to 11, whereas n_2 would be 61, five and a half times greater, with per-priority effective bandwidths.

In Example 4, we consider the case where the higher-priority queue contains the superposition of CBR ATM connections. We model this superposition as a Poisson process, where each arrival offers one unit of work. (If the ATM connections have not been jittered, the Poisson assumption is conservative.) Let the priority 2 connections be the same as in Example 1: each connection is a two-state on-off MMPP with mean rate 0.01, fraction of time on 0.1, and mean burst size 20. Let the performance parameters be $b_1 = 100$, $b_2 = 5{,}000$, $p_1 = 10^{-9}$, and $p_2 = 10^{-6}$. The admissible sets are given in Fig. 8. Here the maximum priority 1 occupancy is 90 percent, which is higher than in the previous examples, and the gain from the per-priority effective bandwidths is rather small, although at the larger priority 1 occupancies we still see some gain. At 80 percent priority 1 occupancy, the priority-insensitive effective bandwidths restrict n_2 to 10, whereas n_2 would be 18 with per-priority effective bandwidths. Overall, the additional complexity may outweigh the benefit in this example. As indicated before, we could then elect to set e_1^k equal to e_1^1 in each priority constraint of a multiple priority system, such as Eq. 4.

Conclusions

Our main conclusion is that to realize the gains from implementing service priorities at network nodes, the connection admission control and dimensioning policies using effective bandwidths should be revised. A given connection should be associated with multiple effective bandwidths: one corresponding to the priority level of the given connection and (potentially) one for each of the lower-level priorities. Geometrically, our findings can be expressed by saying that the triangular FIFO admissible set A in Fig. 4 should be replaced by the larger trapezoidal priority admissible set $A + B + C$ in Fig. 4.

It should be noted that for some service types, distinct effective bandwidths for all lower priorities may yield only modest efficiency gains, in which case, in order to reduce complexity a given priority i type j connection would have the same value for the effective bandwidth e_{ij}^k for different priority levels k.

We have indicated one way that the per-priority effective bandwidths can be determined in the fourth section, namely by a Gaussian approximation for a bufferless model. Three other methods are described in [8], one being a modification of the now-familiar large-buffer asymptotics associated with the FIFO discipline. The general approach also allows effective bandwidths to be obtained in other ways.

Constructing the new admissible set with priorities shows the advantage of priorities when lower-priority classes have substantially looser performance criteria, because we can see that the admissible set is much larger than without priorities.

References

[1] K. W. Ross, *Multiservice Loss Models for Broadband Telecommunications Networks*, London: Springer-Verlag, 1995.

[2] G. L. Choudhury, K. K. Leung and W. Whitt, "An inversion algorithm to compute blocking probabilities in loss networks with state-dependent rates," *IEEE/ACM Trans. Networking*, vol. 3, 1995, pp. 585–601.

[3] G. L. Choudhury, K. K. Leung, and W. Whitt, "Efficiently providing multiple grades of service with protection against overloads in shared resources," *AT&T Tech. J.*, vol. 74, 1995, pp. 50–63.

[4] C. S. Chang and J. A. Thomas, "Effective bandwidths in high-speed digital networks," *IEEE JSAC*, vol. 13, 1995, pp. 1091–1100.

[5] G. de Veciana, G. Kesidis, and J. Walrand, "Resource management in wide-area ATM networks using effective bandwidths," *IEEE JSAC*, vol. 13, 1995, pp. 1081–90.

[6] F. P. Kelly, "Notes on effective bandwidths," *Stochastic Networks*, F. P. Kelly, S. Zachary, I. Ziedins, Eds., Oxford: Clarendon Press, 1996, pp. 141–68.

[7] G. L. Choudhury, D. M. Lucantoni and W. Whitt, "Squeezing the most out of ATM," *IEEE Trans. Commun.*, vol. 44, 1996, pp. 203–17.

[8] A. W. Berger and W. Whitt, "Effective bandwidths with priorities," to appear, *IEEE/ACM Trans. Networking*, 1998.

[9] G. de Veciana and G. Kesidis, "Bandwidth allocation for multiple qualities of service using generalized processor sharing," *IEEE Trans. Info. Theory*, vol. 42, 1996, pp. 268–72.

[10] Z.-L. Zhang, "Large deviations and the generalized processor sharing scheduling for two-queue systems," *Queueing Systems*, vol. 26, 1997, pp. 229–54.

[11] A. I. Elwalid and D. Mitra, "Analysis, approximations and admission control of a multi-service multiplexing system with priorities," *Proc. IEEE INFOCOM '95*, 1995, pp. 463–72.

[12] V. G. Kulkarni and N. Gautam, "Admission control of multi-class traffic with service priorities in high-speed networks," *Queueing Systems*, vol. 27, 1997, pp. 79–97.

[13] A. Feldmann *et al.*, "The changing nature of network traffic: scaling phenomena," *Comp. Commun. Rev.*, vol. 28, no. 2, Apr. 1998, pp. 5–29.

Biographies

Arthur W. Berger [SM '97] (awberger@att.com) received a Ph.D. in applied mathematics from Harvard University in 1983. He then joined AT&T Bell Laboratories, and, in 1996, AT&T Laboratories. He is now with Lucent Technologies Bell Laboratories. He has worked in the areas of network planning, performance analysis of telecommunication switching systems, and B-ISDN/ATM on the topics of congestion controls and traffic engineering. On the latter topics he has been active in ITU Study Groups 2 and 13 and in the ATM Forum. His research interests are in traffic and congestion controls and traffic engineering for high-speed communication networks. He is a member of ACM SIGCOMM.

Ward Whitt (wow@research.lucent.com) received a Ph.D. in operations research from Cornell University in 1969. He was on the faculty of Stanford University in 1968–1969 and Yale University in 1969–1977. He joined AT&T Bell Laboratories in 1977, AT&T Laboratories in 1996, and then in June 1998 joined Lucent Technologies. He is currently a member of the Network Mathematics Research Department in the Networking and Distributed Systems Research Laboratory of AT&T Laboratories. His research has primarily been in queuing theory and its applications to telecommunication systems. He is a member of the Institute for Operations Research and Management Sciences and the Institute of Mathematical Statistics.