CHAPTER 4

Connection Admission Control

An ATM connection traverses a set of switching nodes along its path. Even within a switching node, a connection may traverse a number of possible congestion (or queuing) points. To set-up a connection on such a path, resources need to be reserved to guarantee the contracted Quality-of-Service (QoS) [see Chapter 2]. Generally, the resources needed are the buffer space and the bandwidth required to serve the connection at a queuing point. The sets of rules (or procedures) which determine the admissibility of a connection in an ATM switch are generally termed as *Connection Admission Control (CAC).*

As discussed in Chapter 2, an ATM connection can carry traffic of a particular service category; e.g., CBR, VBR, ABR or UBR. Further more, each of these service categories may possess differing QoS objectives. Thus, the CAC rules are likely to be different for each service category. Each service category also has specific traffic descriptors associated with it and different traffic models are necessary for each service category

To verify the admissibility of a connection, the Connection Admission Control follows the general procedures outlined below to set-up a connection:

1. Map the traffic descriptors associated with a connection onto a traffic model;
2. Use this traffic model with an appropriate queuing model for each congestion point, to estimate whether there are enough system resources to admit the connection in order to guarantee the QoS at every congestion (or queuing) point.
3. Allocate resources if the connection is accepted.

Depending on the traffic models used, the CAC procedures can be too conservative by over allocating the resources. This reduces the *statistical gains* (defined later in the chapter) that can be obtained. An efficient CAC is the one which produces maximum amount of statistical gain at a given congestion point without violating the QoS. The efficiency of the CAC thus depends on how closely the two steps above model reality. Both the traffic and queuing models are well researched and widely published in the literature.

It should be noted that the CAC algorithms cannot be computationally intensive as these are executed in real-time by the ATM switches. This execution directly affects the call set-up rate and the set-up delay as the CAC algorithm is executed on every call set-up. Since connections are set-up and torn down in real-time, handling a high call set-up rate is crucial for efficiently supporting Switched Virtual Circuits (SVCs).

There may be other techniques to allocate bandwidth based on heuristics, or long term measurements. These types of connection admission control procedures cannot guarantee the QoS and do not allow for safe overbooking of the resources (see section Tuning the Connection Admission Control).

The CAC algorithms are not specified by the ATM Forum or the ITU-T, because each switch architecture, queuing and scheduling implementations may be more suited for a specific type of CAC implementation. Other limitations, such as processing capacity or buffer size may dictate the use of a specific CAC implementation. It is not necessary to have the same CAC function on every switch or even on every queuing point within a switch to achieve end-to-end QoS guarantees.

This chapter details some of the procedures for connection admission control that can be adopted by ATM switches. Comprehension of these algorithms is not necessary to understand the remainder of this book. First, we look at the traffic, queuing models and CAC functions for the CBR traffic. The CAC methods for CBR traffic are divided into two areas: the methods that neglect CDV and the methods that account for the CDV. Then the CAC functions for VBR are discussed. There are two distinct CAC methods for VBR traffic: methods that consider each connection independently and the methods that consider all the connections together. The former method is also called “Effective bandwidth” method. The CAC extensions to multi-class traffic and the effect of CDV are also discussed. The CAC procedures based on measurement techniques are presented. The CAC rules for ABR and UBR services are also discussed. Finally, different ways of tuning the CAC function are described.

STATISTICAL GAIN

Since a connection's maximum possible data rate is the Peak Cell Rate (PCR), ignoring the jitter (or CDV), a CAC need not allocate more than PCR to a connection. However, since connections do not continuously send data, it is possible to allocate less for some traffic classes when many connections are statistically multiplexed at a congestion point. This means that "statistical gain" is possible and more connections can be admitted when compared to peak rate allocation. Thus, the term "statistical gain" can be defined as:

$$Stastical\,Gain = \frac{Number\,Connections\,admitted\,with\,Statistical\,Multiplexing}{Number\,of\,Connections\,admitted\,with\,peak\,rate\,allocation}$$

An efficient CAC should try to achieve as much statistical gain as possible without entering a congested condition causing cell loss on connections. The gain (equivalently, the amount of statistical multiplexing that can be achieved), is generally a function of buffer size, traffic characteristics and QoS objectives of connections that are being multiplexed. For a given traffic and QoS parameters, larger statistical gain can be obtained with larger buffers. Thus, the congestion at a queuing point [RMV96] can be divided into two parts:

1. Cell-scale congestion that occurs in a small buffer,
2. Burst-scale congestion that occurs typically in a large buffer.

The CBR and real-time VBR (rt-VBR) service categories have nodal delay requirements. That is, the nodal delay experienced by cells of these services should not be more than a given value D (e.g., 250μs) with a given quantile Q (e.g., 10^{-10}), i.e., $P(Nodal\ Delay > D) \leq Q$. This delay requirement forces the buffer sizes to be very small and the cell-scale congestion will be prevalent for these services. Therefore, it is difficult to achieve large statistical multiplexing gain for CBR and rt-VBR services. For nrt-VBR, the ATM switches provide larger buffers to absorb the bursts and the burst-scale congestion will occur frequently for these services. Therefore, it is generally possible to achieve large statistical gain for nrt-VBR services.

CAC FOR CBR TRAFFIC

A pure CBR traffic source emits a cell periodically at every *1/PCR* seconds. Ignoring the CDV and the cell scale congestion, a simple rule of CAC is to assign the *PCR* as the bandwidth required for each CBR connection and admit connections such that,

$$\sum_i PCR_i \leq Link\,Capacity \qquad (4.1)$$

This is termed as CAC based on "peak rate allocation" [see Chapter 3]. As ATM connections traverse through various queuing points, the periodicity of a pure CBR source will be lost. The cells may clump together or disperse due to buffering and the effects of other intervening traffic. This is also called "jitter" or "Cell Delay Variation (CDV)". Thus, each cell may experience a different transfer delay. When cells are clumped together, the effective peak rate will be higher than the source actual peak rate. Due to the presence of this CDV, this simple CAC rule is not sufficient to ensure that the Cell Loss Ratio (CLR) of the admitted connections is within the prescribed limits. Even without this jitter, the simultaneous arrival of cells due to the multiplexing of periodic cell streams can also cause cell loss. This buffer overflow can occur typically in a small buffer. In general, there are two approaches to account for CDV [RMV96]: 1) negligible CDV methods, 2) non-negligible CDV methods.

Negligible CDV Methods

These methods do not account for the CDV directly and assume that the connections have negligible CDV. One such method is to model the multiplexer queue as an *M/D/1* queue. Given the buffer size *B*, the link capacity *C* and the peak cell rate of the connection PCR_i, determine a load ρ such that the probability of queue length exceeding *B* is less than ε, where ε is a small number such as 10^{-10}. The CAC rule is to admit the connections until,

$$\sum_i PCR_i \le \rho \times Link\,Capacity \qquad (4.2)$$

The second method is to use a *nD/D/1* multiplexer model [DRS91, RV91, RMV96, FLV94] that serves *n* identical periodic CBR cell streams. This method produces less conservative results than an *M/D/1* model for high loads (>80%). Asymptotic methods in references [DRS91, RMV96, and FLV94] are not computationally intensive and serve as good approximations. For example in [FLV94], the *M/D/1* approximation is given as:

$$P(Buffer\,Length > x) \approx -\frac{1-\rho}{\ln(\rho)}\exp\left(-x\left(1-\rho-\ln(\rho)\right)\right) \qquad (4.3)$$

The *nD/D/1* approximation based on Brownian Bridge approximation is given by [FLV94]:

$$P(Buffer\,Length > x) \approx -\frac{1-\rho}{\ln(\rho)}\exp\left(-x\left(\frac{2x}{n}+1-\rho-\ln(\rho)\right)\right) \qquad (4.4)$$

Here, it is assumed that the service duration is one cell slot and the units of buffer length and *x* are in number of cells. The following graph compares these two methods for various buffer sizes:

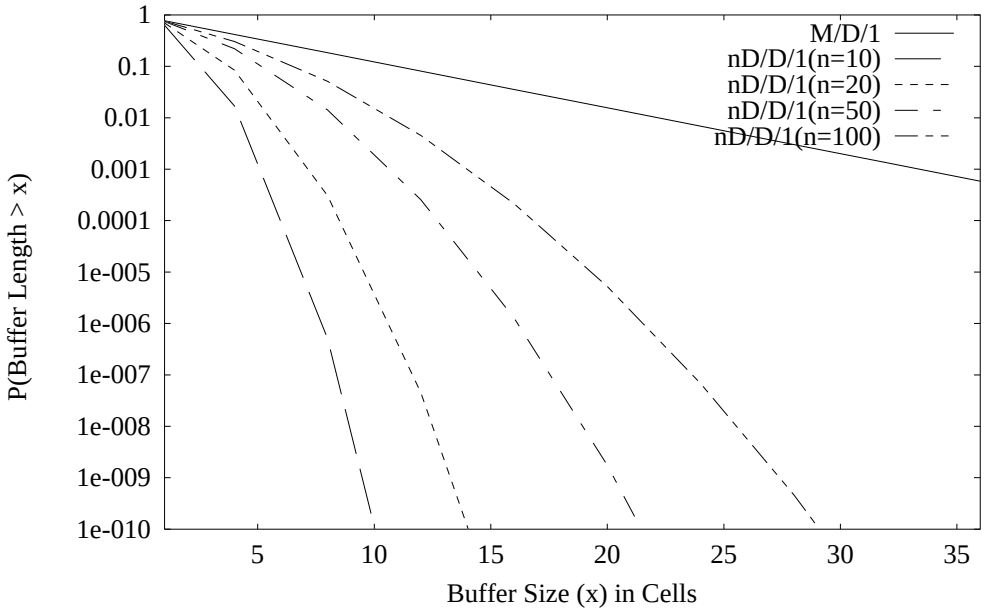

Figure 4.1 Cell Loss Probability vs. Buffer Size

As indicated by Figure 4.1, *M/D/1* model is conservative and the *nD/D/1* model can obtain better capacity. For large *n*, both models give nearly the same result as the system is well approximated by the Poisson arrivals.

The above models assume a homogeneous system of sources, i.e., all sources have identical *PCR*. In reality, the sources will have different *PCR* requirements. Therefore, models like $\Sigma D_i/D/1$ [RMV96], $\Sigma n_k D_k/D/1$ [RV91] are useful for this purpose. Another way is to map the heterogeneous model into an equivalent homogeneous system. One such example is shown in [FLV94] by Fiche et.al.

Non-Negligible CDV Methods

The negligible CDV method is typical of the multiplexing of pure CBR sources. The CDV produced may not be negligible when CBR traffic streams are multiplexed with other traffic streams such as rt-VBR. In this case, the traffic stream is bursty and buffering should be sufficient to absorb the burst scale congestion. Each ATM connection is generally policed at the network edges using the Generic Cell Rate Algorithms (GCRA) [see Chapter 3]. For CBR connections arriving on a link with a link rate (*LR*), the peak rate policing is enforced with *GCRA(1/PCR, CDVT)*. The worst case output traffic pattern of such a policing function

will be a periodic on-off process with a maximum burst size $BS=1+\lfloor CDVT/(T-\delta) \rfloor$ where $T = 1/PCR$ and $\delta = 1/LR$.

Therefore, to account for the CDV, in addition to Equation (4.1), another constraint can be placed on the buffer such that $\sum BS_i \leq B$, where B is the buffer size of the multiplexer. This worst case estimate assumes that no cell loss occurs with the simultaneous arrivals of bursts from all the connections (also called "lossless multiplexing"). However, this constraint can be very pessimistic. Another way is to consider the on-off process of the leaky bucket model and map it to an equivalent VBR traffic with parameters $SCR_{VBR} = PCR_{CBR}$, $PCR_{VBR} = LR_{CBR}$ and $MBS = 1+\lfloor CDVT/(T-\delta) \rfloor$. Then, the same CAC as for the VBR traffic can be applied. Here, PCR_{CBR} is the peak cell rate of CBR connection; LR_{CBR} is the link rate of the CBR connection. The CAC methods for the VBR traffic are described in the next section.

CAC FOR VBR TRAFFIC

As noted in Chapter 2, the Variable Bit Rate (VBR) traffic is generally characterized by an average rate (SCR), Peak Cell Rate (PCR), Cell Delay Variation Tolerance (CDVT) and Maximum Burst Size (MBS). The QoS specifications for rt-VBR traffic are cell loss ratio (CLR), delay while nrt-VBR traffic has only the cell loss.

The ratio SCR/PCR defines the *burstiness* of the VBR traffic and has a strong impact on the statistical gain. If the burstiness (or SCR/PCR)<< 1, a CAC based on peak-rate allocation will be very inefficient and conservative. It is shown [WK90] that the link utilization can go as low as 5% when SCR/PCR is small. Even if the buffer for VBR traffic is very small, peak-rate allocation is unnecessary. It is possible to admit connections such that,

$$\sum_i BW_i \leq LinkCapacity \qquad (4.5)$$

where, $SCR_i \leq BW_i \leq PCR_i$. While peak-rate allocation does not provide any statistical gain, Equation (4.5) provides the statistical gain. The statistical gain at a queuing point can then be defined as the ratio $\sum PCR_i / \sum BW_i$. The BW_i is also called the "Equivalent Bandwidth (EBW)" or "Virtual Bandwidth (VBW)" of a connection.

In the case of small or no buffer scenario, *Rate Envelope Multiplexing (REM)* [RMV96] technique can be used for the CAC. That is, the connections are admitted such that the total aggregate arrival rate (rate envelope) of the connections is less than the link capacity with a high probability. On the other hand, if there is a very large buffer ($\rightarrow \infty$) available for the VBR traffic, it would absorb the bursts and one need not allocate more than SCR for each VBR connection. In practice, the buffer is finite and the bandwidth allocation falls in between that of SCR and PCR. This method is called *Rate Sharing (RS)* [RMV96] technique.

Rate Envelope Multiplexing (REM)

The *Rate Envelope Multiplexing (REM)* method [RMV96] assumes that there is little or no buffering available to the VBR traffic. Thus, it is also called the *zero buffer* or *bufferless* approximation. This method is well suited for real-time traffic due to the assumption of small buffers and models the cell-scale congestion well. By this method, the connections are admitted so that the aggregate arrival rate *AR* (i.e., rate envelope) of the connections is less than the link capacity *C* of the queuing point with large probability. The cell loss ratio (*CLR*) can be estimated as:

$$CLR = \frac{E\left\{(AR - C)^+\right\}}{E\{AR\}} \tag{4.6}$$

Here, the CLR is defined as the ratio of amount of work lost to the amount of work arrived. It is only dependent on the connection parameters and not on the queuing behavior of the system. The operator $(.)^+$ only takes into account the positive differences (i.e., when $AR > C$) and is zero when $AR < C$. For call admission purposes, the aggregate rate *AR* can be measured in either real-time or estimated from the traffic models. Once *AR* is known, the *CLR* is estimated using Equation (4.6) before admitting a new connection. The connection can be accepted if the resulting CLR is lower than the objective. For example, when *N* identical on-off sources with peak rate *PCR* and mean rate *SCR* are multiplexed on a link of capacity *C*, the *CLR* is given by [RMV96]:

$$CLR = \sum_{i.PCR>C} (i.PCR - C)\binom{N}{i}\left(\frac{SCR}{PCR}\right)^i\left(1 - \frac{SCR}{PCR}\right)^{N-i} \Big/ N.SCR \tag{4.7}$$

Alternately, one can use the equivalent bandwidth approach. This is described in Section “Effective Bandwidths”, later in the chapter.

Rate Sharing (RS)

The REM method relies on the fact that the total aggregate input rate does not exceed the link capacity or the probability of exceeding is very small. Otherwise, buffering is needed to absorb any rate mismatch. This is especially true for bursty traffic, whose average rate is small compared to their peak rate. For example, for the VBR traffic with *SCR* << *PCR*, there could be a small number of connections simultaneously arriving at *PCR* and the total such instantaneous rate can be much larger than the link capacity. Thus, to guarantee a certain QoS to the connections, such as *CLR*, a buffer is needed for temporary storage of the traffic and the link capacity is shared among the contending connections. Thus, the queuing models should also be incorporated into the connection admission rules.

There are two approaches to do this:

1. Consider all the traffic streams that are being multiplexed together and estimate the cell loss probability to obtain maximum possible statistical gain;
2. Consider each connection independently from other traffic streams and estimate bandwidth requirements of the connection, so as to guarantee a given QoS. This method is also referred to as the effective bandwidth approach.

The effective bandwidth approach is explained in detail in the next section. The first approach considers the connections that are already admitted plus the one being admitted and checks whether the new connection would violate the QoS guarantees of any of the connections. The new connection will be accepted if the QoS specifications are met. Connection admission control based on such statistical multiplexing approach is discussed in this section.

In [BC92], a large class of VBR sources and their superposition is taken into account. A point-process belonging to the class of discrete-time batch Markovian arrival processes (D-BMAP) is proposed. Buffer occupancy and cell loss probabilities of a statistical multiplexer are derived based on this process using a matrix-analytical approach.

Buffet and Duffield [BD94] developed exponential upper bounds for the queue distribution of FCFS (First-Come-First-Serve) queue, which is fed by superposition of homogeneous Markovian on-off sources. Using the theory of Martingales, a bound of the form $P[\text{Queue Length} \geq B] \leq cy^{-B}$ for any $B \geq 1$ where $c<1$ and $y>1$ is given explicitly as a function of parameters of the model. The model assumes that L independent markovian sources are multiplexed, each with mean silence length of $1/a$ units, and the mean burst length of $1/d$ units. Let the service rate of the multiplexer be σL. Define $k=(1-a-d)$. Then, the explicit bound for the cell loss probability is given by:

$$P[Q \geq B] \leq \frac{a(1-a-\sigma k)}{d(a+\sigma k)} \left[\frac{(1-\sigma)(a+\sigma k)}{\sigma(1-a-\sigma k)} \right]^{B} \left[\frac{1}{(a+d)} \left(\frac{d}{1-\sigma} \right)^{1-\sigma} \left(\frac{a}{\sigma} \right)^{\sigma} \right]^{L} \tag{4.8}$$

This method is extended to superposition of heterogeneous sources by Duffield [Duf92]. An *MMDP/D/1/K* queuing system is used in [YT95] to estimate the cell loss probability of an ATM multiplexer loaded with homogeneous on-off sources. As per the *Markov Modulated Deterministic Process (MMDP)*, a source can be in one of the two states: *on* and *off*. When the source is on, it generates a stream of cells that are equally spaced at a fixed rate, called the peak rate. When it is off, the source generates no cells. Both on and off periods are distributed as independent exponential random variables. Exact and approximate methods were developed to estimate the cell loss probability.

Diffusion process approximation of a statistical multiplexer is considered by Ren and Kobayashi [RK94]. The sources are modeled as Markov Modulated Rate Process (MMRP)

and a diffusion process approximation for the superposition of such rate processes is developed in [RK94]. This model is not restricted to the exponential distribution for the duration of each state.

There are many more traffic as well as queuing models proposed in the literature for the CAC purposes and it is difficult to list all of them here. However, it should be observed that all these methods estimate the tail probabilities i.e., *P(Queue Length > Buffer)* as a function of *N*, the number of connections. Although these methods may achieve better statistical gain than the methods based on *Effective Bandwidths* (see next section), the state dependence on the number of connections (*N*) make them generally difficult to implement in real switches. These methods tend to be computationally intensive and such processing power may be limited in the switches. The CAC function should be able to dynamically add or remove connections in real-time. Therefore, methods that translate the source traffic parameters into one number called the Equivalent or Effective Bandwidth have become very popular. Many ATM network-engineering tools use the effective bandwidths approach to estimate the required bandwidth of connections.

Effective Bandwidths

The effective bandwidth approach views each connection in isolation, as if it were present alone at the congestion point. This model maps each connection's traffic parameters into a real number EBW_i, called the *Equivalent Bandwidth* or *Effective Bandwidth* of the connection such that the QoS constraints are satisfied. Thus, the effective bandwidth is derived as a source property and with this mapping, the CAC rule becomes very simple:

$$\sum EBW_i \leq Link\,Capacity.$$

That is, a connection is admitted if there is available spare capacity or else it is rejected. For a connection with an average rate SCR_i and peak rate as PCR_i, the effective bandwidth is a number between the SCR_i and PCR_i. That is, $SCR_i \leq EBW_i \leq PCR_i$. The value of effective bandwidth depends upon the statistical properties of the connection being admitted as well as on the queuing properties of the congestion point under consideration. In general, for a given connection, it is intuitive that the effective bandwidth will be close to the peak rate for very small buffers and close to average rate for very large buffers.

The main advantages with the effective bandwidth method are:

1. *Additive Property*: Effective bandwidths are additive, i.e., the total effective bandwidth needed for *N* connections equals to the sum of effective bandwidth of each connection
2. *Independence Property*: Effective bandwidth for a given connection is only a function of that connection's parameters.

The additive property of the effective bandwidth method makes it widely accepted and used in ATM technology. However, it should be noted that the due to the independence property, the effective bandwidth method could be far more conservative than a method which considers the true statistical multiplexing (i.e., the method which considers the presence of other connections). The main reason for this is that the actual bandwidth needed to serve *N* connections could be far less than the sum of bandwidths of *N* connections. In the world of ATM, connections are set-up and torn down dynamically. Thus, with the effective bandwidth's method, the CAC function can add (or subtract) the effective bandwidth of the connection which is being set-up (or torn down) from the total effective bandwidth. This is not easily possible with any method which does not have the independence property. In the following sections, effective bandwidths for rate envelope multiplexing and rate sharing are presented.

Effective Bandwidths for Rate Envelope Multiplexing

The CLR for rate envelope multiplexing is estimated by Equation (4.6). Kelly [Kel91] developed the effective bandwidths for a heterogeneous system of sources producing an aggregate load

$$AR = \sum_{j=1}^{J} \sum_{i=1}^{n_j} AR_{ji} \tag{4.9}$$

Here, AR_j is the load produced by class j and n_j is the number of sources in class j. Using a Chernoff bound, Kelly [Kel91] showed that,

$$\log P(AR \geq C) \approx \inf_{s} \left[\sum_{j=1}^{J} \left(n_j M_j(s) - sC \right) \right] \tag{4.10}$$

where,

$$M_j(s) = \log E\left[e^{sAR_{ji}} \right] \tag{4.11}$$

is the logarithmic moment generating function of the random variable AR_{ji}. With s attaining an infimum at s^*, the call acceptance region becomes:

$$A(n^*) = \left\{ n : \sum_{j=1}^{J} \alpha_j^* n_j + \frac{\gamma}{s^*} \leq C \right\} \tag{4.12}$$

Here, $\alpha_j^* = M_j(s^*)/s^*$ is the *effective bandwidth* of source *j*, and $\log P\{AR \le C\} \le -\gamma$. For on-off sources with peak rate PCR_j and average rate SCR_j, the moment generating function is given by,

$$M_j(s) = \log\left(1 + \frac{SCR_j}{PCR_j}\left(e^{sPCR_j} - 1\right)\right) \qquad (4.13)$$

Effective Bandwidths for Rate Sharing

The rate-sharing technique uses a buffer to absorb momentary bursts of data. That is, the aggregate arrival rate can momentarily exceed the service capacity. The buffer can be dimensioned to allow small loss probabilities. The effective bandwidth method should consider the cell loss ratio and the buffer size. The techniques of effective bandwidths for rate sharing can be divided into lossless and loss tolerant models. As the name implies, lossless models do not allow cell loss to occur and thus CLR is not explicitly taken into account.

Kelly [Kel91] developed effective bandwidth for a *M/G/1* model for lossless performance. Here, it is assumed that bursts from a source of class *j* arrive in a Poisson stream of rate r_j. The bursts are assumed to have a length distribution G_j. Let μ_j, σ_j^2 be the mean and variance of G_j and L the buffer size. Then the effective bandwidth of a source of type *j* is found to be:

$$\alpha_j = r_j\left[\mu_j + \frac{1}{2L}\left(\mu_j^2 + \sigma_j^2\right)\right] \qquad (4.14)$$

For a source which is policed by a leaky bucket algorithm, Elwalid [EMW95] developed the effective bandwidth. The departure process of such a policing function is assumed to be on-off and periodic. Let B_T be the token buffer size of such regulator. Let SCR_j, PCR_j be the average and peak cell rates of connection *j*. The maximum burst size (MBS) allowed by the regulator will be $MBS_j = B_T\lfloor PCR_j/(PCR_j - SCR_j)\rfloor$. The effective bandwidth of such a connection for lossless performance is given by:

$$\alpha_j = \begin{cases} \dfrac{PCR_j}{1+\dfrac{B(PCR_j - SCR_j)}{B_T C}}, & \text{if } SCR_j \le \dfrac{B_T}{B/C} \\ SCR_j & \text{if } \dfrac{B_T}{B/C} \le SCR_j \le PCR_j \end{cases} \tag{4.15}$$

The loss probability is considered in realistic models through the *asymptotic slope* of the queue length distribution. There is a plethora of research on this topic. In general, for a given buffer size *B*, the queue length tail probabilities are asymptotically exponential and given by:

$$P(\text{Queue Length} \ge B) \approx e^{-f(c_i)B} \tag{4.16}$$

The function *f()* is determined from the statistical properties of the traffic stream. The term c_i is the effective service rate or effective bandwidth needed to serve the connection in order to guarantee a given *CLR*, i.e., *P(Queue Length > B)* $\le$ *CLR*. From these two equations, the effective bandwidth of a connection can be determined as:

$$c_i = f^{-1}\left(-(\log CLR)/B\right) \tag{4.17}$$

Let $(\log CLR)/B = \zeta \in [-\infty, 0]$. For on-off sources with exponentially distributed 'on' and 'off' periods, Gibbens and Hunt [GH91] derived the effective bandwidths of a source. Let a source mean "on" period be $1/\mu_i$ and mean "off" period be $1/\lambda_i$. When the source is "on", it is assumed to produce information at a constant rate γ_i. Then, the effective bandwidth c_i of the source is given by Gibbens and Hunt [GH91] as:

$$c_i = \frac{\left(\zeta\gamma_i + \mu_i + \lambda_i\right) - \sqrt{\left(\zeta\gamma_i + \mu_i - \lambda_i\right)^2 + 4\lambda_i\mu_i}}{2\zeta} \tag{4.18}$$

Equation (4.18) implies that for large *B*, $\zeta \to 0$ and c_i equals to the mean rate of the source $\lambda_i\gamma_i/(\lambda_i + \mu_i)$. For a small buffer *B*, $\zeta \to -\infty$ and the effective bandwidth of the source will be $c_i = \gamma_i$, the peak information rate. For a connection generating traffic using the on-off model with traffic descriptors of SCR, PCR, CLR and ABS (Average Burst Size), the

above parameters can be mapped as: $\mu_i = PCR/ABS$, $\lambda_i = \mu_i.SCR_i/(PCR-SCR)$, $\gamma_i = PCR$. Note that for VBR connections, *MBS* and not *ABS* is indicated in the traffic descriptor. Therefore, in order to use this function as a CAC for VBR traffic, another mapping will be needed which translates the ABS into an MBS. This is not discussed here.

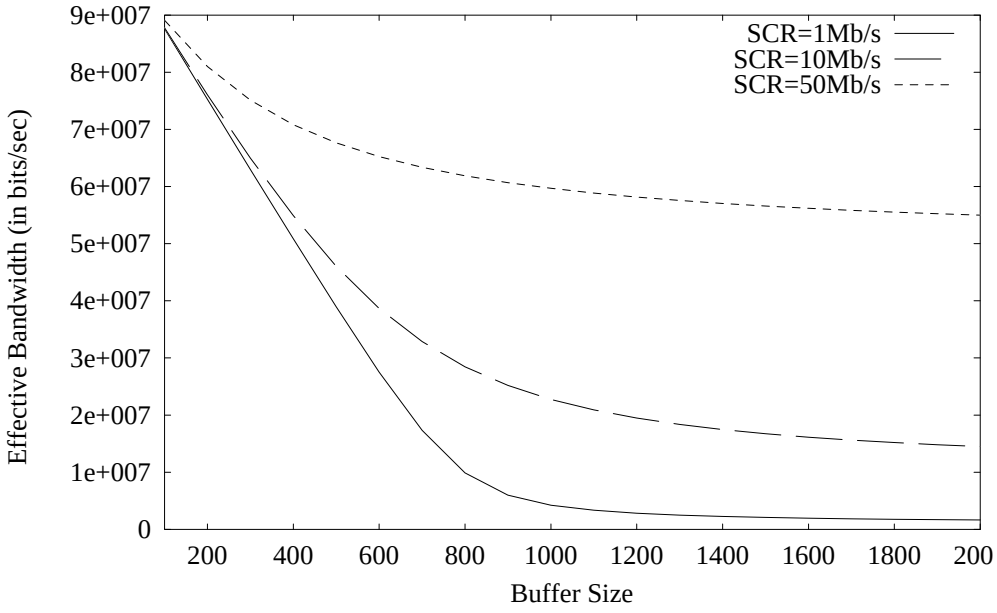

Figure 4.1 Effective Bandwidth as a function of buffer size

Using equation (4.18), Figure 4.2 shows the effective bandwidth as a function of buffer size for various traffic parameters. The connection with a PCR = 100 Mb/s, CLR = 10^{-7} and ABS = 50 cells is chosen for this graph. The effective bandwidth as a function of buffer size is plotted for various values of SCR. It can be seen clearly from this figure that the effective bandwidth is close to PCR for small buffers and tends to approach SCR as buffer size is increased. Figure 4.3 shows the amount of statistical gain that can be achieved for the same connections.

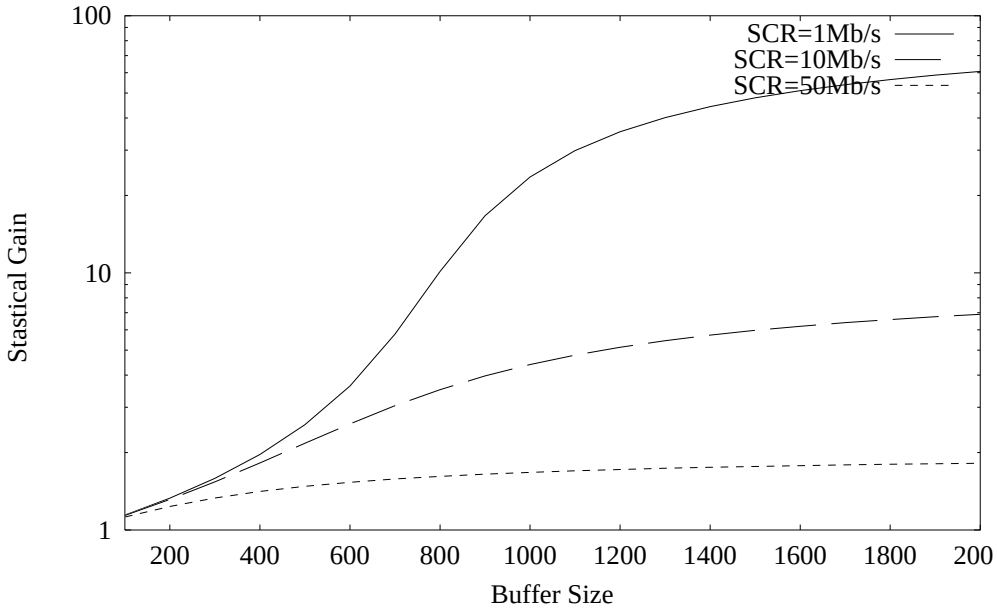

Figure 4.2 Statistical Gain as a function of Buffer Size

The result of Gibbens and Hunt [GH91] is extended in [EM93]. Elwalid and Mitra [EM93] obtain two sets of results: one for a statistical multiplexing with general Markov-modulated fluid sources and the other for queues in which traffic sources are Markov-modulated Poisson or phase renewal process. For large B, an approximate measure of effective bandwidth is given by Courcoubetis et.al [CFW94] as:

$$c_i = m_i + \frac{\delta \xi_i}{2B} \tag{4.19}$$

where, m_i is the mean rate of the source, δ = *-log (CLR)* and ξ_i is called the index of dispersion. For an on/off Markov fluid model with a mean "on" period $1/\mu_i$, a mean "off" period $1/\lambda_i$ and peak information rate γ_i, the effective bandwidth will be:

$$c_i = \frac{\lambda_i . \gamma_i}{\lambda_i + \mu_i} + \frac{\delta \lambda_i \mu_i \gamma_i^{\,2}}{B\left(\lambda_i + \mu_i\right)^3} \tag{4.20}$$

To see how these effective bandwidth approximations compare, three methods are considered below. These are: 1) Gibbens and Hunt method given by equation (4.18), 2) Courcoubetis et.al method given by equation (4.20) and 3) Buffet and Duffield method given by equation (4.8). Note that equation (4.8) does not give any explicit effective bandwidth. Instead, the equation is used to calculate the effective service rate required in serving a single connection so that the loss probability is less than a given value. Figure 4.4 shows such a comparison for connections with PCR = 100Mb/s, CLR=10^{-7}, Line Rate = 150 Mb/s, Average Burst Size (ABS) = 50 cells. Although this figure is illustrated as a means of comparison, it should be kept in mind that the degree of conservatism between methods changes with traffic parameters. Although not illustrated, for example for large ABS, equation (4.20) can be more conservative than others for small buffer sizes.

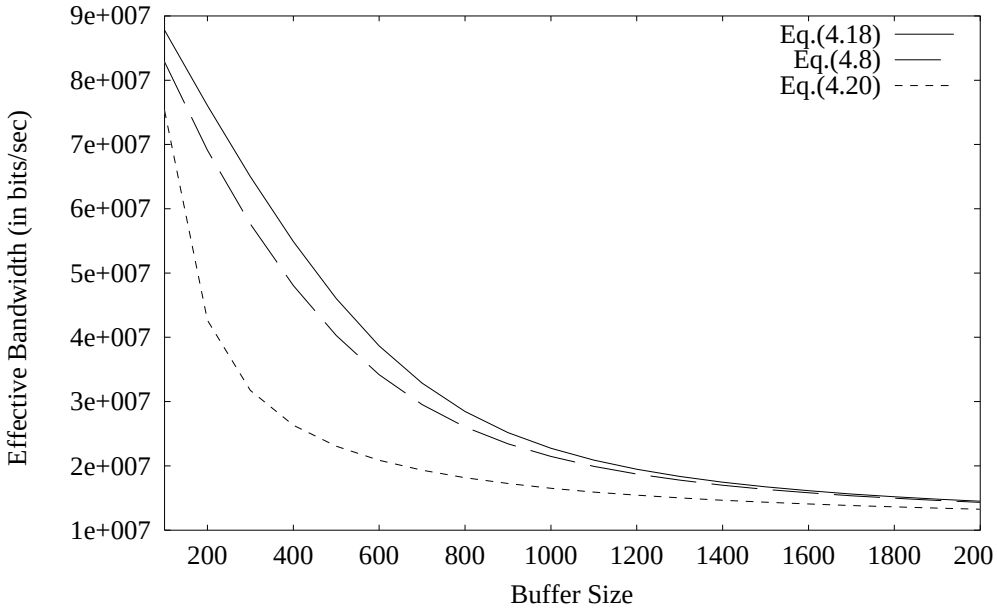

Figure 4.3 Comparison of Effective Bandwidth function

A fluid-flow approximation is used by Guérin et.al [GAN91] in deriving the effective bandwidth for a single two-state Markov source. Let the peak rate of the source be PCR_i, utilization ρ and mean burst period b. Let the queue size be B and $\alpha = \ln(1/CLR)$. Then the effective bandwidth of the source is approximated by:

$$c_i = \frac{\alpha b(1-\rho)PCR_i - B + \sqrt{[\alpha b(1-\rho)PCR_i - B]^2 + 4B\alpha b\rho(1-\rho)PCR_i}}{2\alpha b(1-\rho)} \quad (4.21)$$

It is possible that Equation (4.21) can be very conservative for the case of large bursts. In general, the effective bandwidth methods are conservative since they do not consider the effect of multiplexing many sources. Therefore, an aggregate stationary bit rate approximation is also used by Guérin et.al [GAN91] to account for this multiplexing. The distribution of this stationary bit rate is assumed to be Gaussian. Thus, when *N* connections are multiplexed, the total capacity needed is approximated as:

$$C_S = m + \alpha'\sigma \tag{4.22}$$

where, *m* is the mean aggregate bit rate $\left(= \sum_{i=1}^{N} SCR_i\right)$, σ is the standard deviation of the aggregate bit rate and $\alpha' = \sqrt{-2\ln(CLR) - \ln(2\pi)}$. The flow and stationary approximations are then combined into one single approximation. That is, the total equivalent capacity needed to multiplex *N* connections is taken as the minimum of stationary and fluid approximations as $\min\left\{m + \alpha'\sigma, \sum_{i=1}^{N} c_i\right\}$.

The important expression to the effective bandwidth approximation is Equation (4.16), where it is considered that queue length tail probabilities are exponential asymptotically. A better approximation can be achieved by introducing more variables to this equation as described in [CLW96]. Here, the waiting time probabilities are generalized as, $P(W > x) \approx \alpha e^{-\eta x}$ as $x \to \infty$, where α is the asymptotic constant and η is the asymptotic decay rate. Introducing more terms can further refine this equation. The actual effective bandwidth lies between this approximation and that of Equation (4.16).

Much of the work presented so far is based on Markovian models for the traffic sources. However, network traffic measurements have shown that the network traffic is self-similar. That is, the traffic has shown both short and long-range dependence in its correlation. Markovian models in general cannot capture this behavior. Measurements also have shown that with increase in buffer capacity, the resulting cell loss is not reduced exponentially but decreases slowly. An asymptotical upper bound to the overflow probability which decreases hyperbolically with buffer size is developed by Tsybakov [TG97]. Norros [Nor94, Nor96] developed an effective bandwidth model for the self-similar traffic, which is given by:

$$C = m + \left(H^H (1-H)^{(1-H)} \sqrt{-2\ln(CLR)}\right)^{1/H} a^{1/(2H)} B^{-(1-H)/H} m^{1/(2H)} \tag{4.23}$$

where, *m* is the mean bit rate of the traffic stream, *a* is the coefficient of variation, *B* is the buffer size, *H* is the Hurst parameter of the stream *(0.5≤H≤1)*, *CLR* is the target cell loss ratio. Note that this equation does not follow the asymptotic exponential queue length distribution.

ADMISSION CONTROL FOR MULTI-CLASS TRAFFIC

Until now, the CAC functions discussed are for single a QoS class. In the real world, the traffic flow consists of multiple QoS classes, where, the services may be partitioned and queued separately. In ATM, the QoS classes can be classified into three general categories: *real-time*, *non real-time* and *best-effort*. Even within a given QoS category, multiple sub-classes can be differentiated depending on the QoS requirement of each subclass. Priority queuing is a simple architecture to provide a multiple QoS (see Chapter 5 on Queuing and Scheduling) structure. An example is shown in Figure 4.5 below. Note that it may be difficult to provide multiple sub-class service differentiation using this platform and the sophisticated structures needed for that purpose (see Chapter 5).

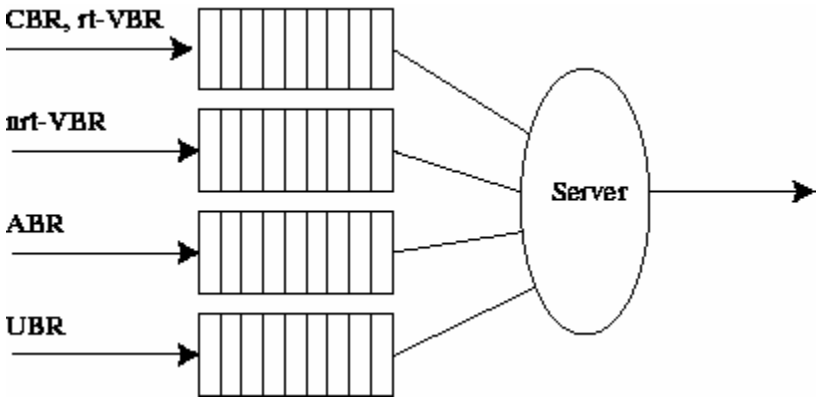

Figure 4.4. A Priority Queuing Structure

To guarantee QoS, a certain amount of bandwidth (or capacity) is reserved for each of the service categories. With effective bandwidth approach, this assignment becomes very simple. Let N_j be the number of sources for class j and let α_j be the effective bandwidth of a source belonging to class j. Let there be K such classes. Then, the CAC for multi-class traffic should check that the total estimated capacity is less than the service rate. That is,

$$\sum_{j=1}^{K} N_j \alpha_j \leq LinkCapacity \qquad (4.24)$$

The existence of effective bandwidths for multi-class Markov fluids and other types of sources that are used to model ATM traffic is shown in [KWC93]. Hsu and Walrand [HW96] proposed an admission control scheme for multi-class ATM traffic in which it is assumed that the real-time traffic is not buffered. In this case, the real-time traffic uses peak

rate allocation as well as provides multiple sub-class QoS. However, the total bandwidth assigned to the real-time traffic for each class is limited to $min(\Sigma PCR, C_k)$, where C_k is the capacity estimated to support the required CLR.

An effective bandwidth vector for a two-priority ATM traffic is presented by Kulkarni et.al in [KGC94]. This model provides a two-subclass QoS using a single buffer. In ATM, the cells of a connection are either marked as CLP=0 or CLP=1. Generally, the QoS guarantees are only specified for CLP=0 traffic. A small threshold is maintained for each queue such that a CLP=1 cell will be dropped if the queue length is greater than this threshold. This way, no specific QoS guarantees are provided for the CLP=1 traffic. One can view the CLP=0 traffic as high priority while the CLP=1 traffic as low priority although both cells are originated from the same source. The equivalent bandwidth function is extended to such two-priority traffic in [KGC94].

EFFECT OF CELL DELAY VARIATION ON CAC

Due to the buffering and contention for service at various congestion points in an ATM network, each cell of a connection will accumulate different transfer delays. To account for this cell delay variation, connections are policed with CDVT parameter at an UPC point. For example, a connection with a peak cell rate as *PCR* and cell delay variation tolerance as *CDVT* is policed by a UPC function *GCRA(PCR,CDVT).* This means, the output of a policing function for a pure CBR source emitting cells at PCR would tend to be bursty within the limits of the specified CDV tolerance. This added burstiness should be considered in the CAC, as this would require more system resources.

The non-negligible CDV methods described earlier in the section of CAC for CBR traffic did not explicitly consider the CDV tolerance. As described in that section, one way is to consider the on-off process of the leaky bucket model and map it to an equivalent VBR traffic with parameters $SCR_{VBR} = PCR_{CBR}$, $PCR_{VBR} = LR_{CBR}$ and $MBS = 1+\lfloor CDVT/(T-\delta) \rfloor$. Then, the same CAC as for the VBR traffic can be applied. Here, PCR_{CBR} is the peak cell rate of CBR connection; LR_{CBR} is the link rate of the CBR connection.

A novel way was devised by Skilros [Skl94] using the CDVT parameter for the CAC. Specifically, the output of the policing function is characterized by a Generalized Geometric (*GGeo*) distribution with two moments, the mean λ and the squared coefficient of variation C_d^2. Let a connection's parameters peak cell rate and CDV tolerance be *PCR, CDVT*. If the connection passes through a multiplexer with a link rate *LR*, then the worst case burst size will be $MBS = 1+\lfloor CDVT/(T-\delta) \rfloor$, where $T = 1/PCR$ and $\delta = 1/LR$. The mean (λ) and coefficient of variation (C_d^2) of the output process of the policing function is then approximated by: $\lambda = 1/PCR$ and $C_d^2 = (MBS-1)(1-(PCR/LR))^2$.

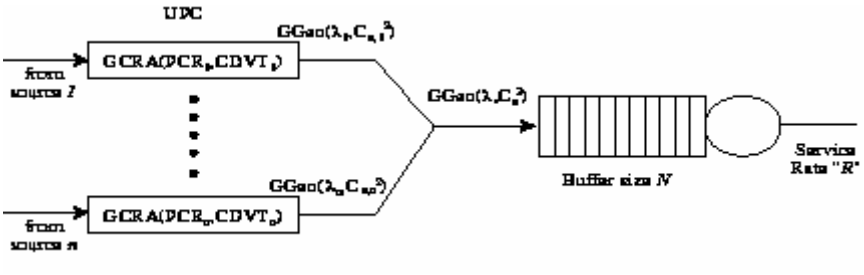

Figure 4.5. Modeling of ATM traffic controls

As shown in Figure 4.6, the individual $GGeo(\lambda_i,C_{di}^2)$ process can then be merged into one single $GGeo(\lambda,C_d^2)$. A maximum entropy solution is formed for the resulting *GGeo/GGeo/1/N* queue, which is used as a CAC function.

CAC BASED ON MEASUREMENTS

The CAC procedures described so far are based on analytical modeling of the behavior of both the traffic sources and queuing structures. In reality, there are varieties of traffic sources and it is impossible to accurately model all of them. Therefore, analytical methods that are used to estimate the resource needs for one given class of traffic sources could severely over or under-estimate the resource requirements for some other class. Hence, CAC procedures based on measurements are becoming increasingly popular. These methods measure certain variables of the system in real-time and attempt to estimate whether the QoS objectives could be maintained if a new connection is admitted.

Courcoubetos et.al [CKW95] proposed one such method. This method monitors the traffic at a buffer and makes estimates of admitting more calls on the fraction of cells lost in that buffer. Given the buffer size *B*, number of connections *N*, service rate *C*, and $F(N,B,C)$ the fraction of cells lost due to buffer overflows, the fraction of cells lost $F(N(1+\varepsilon),B,C)$ is estimated. Here, $N(1+\varepsilon)$ represents the additional calls that are admitted. The function *F* is represented in terms of the buffer overflow probability φ in a busy cycle. For large B, the function $\varphi(N(1+\varepsilon),B,C)$ is approximated as $\varphi(N,B,C/(1+\varepsilon))$. Since F is expressed in terms of φ, it is assumed that $F(N(1+\varepsilon),B,C)$ can also be approximated as $F(N,B,C/(1+\varepsilon))$. The problem now simplifies to estimating the function $F(N,B,C/(1+\varepsilon))$. Since the measured probabilities are very small (of the order of 10^{-7} or less), the function *F* is estimated by using three smaller virtual buffers.

A scheme, which relies on aggregate statistics and cell loss measurements for ongoing traffic, is proposed by in [ZT97]. It is assumed that new calls are considered as CBR transmitting at their peak cell rate for certain warm-up period. This implies that, if the warm-up period is very long, Peak Rate Allocation will be used. The system is adaptive and when the measured cell loss rate is higher than the required rate, the warm-up period is increased. The

cell loss ratio is estimated using Reich's approach, which does not assume any specific model for the traffic. The method in [DJM97] uses both declared traffic parameters and traffic measurements to estimate the aggregate equivalent bandwidth required by the connections. The optimization framework uses a linear Kalman filter for the estimation. It takes into account the connection level dynamics and provides information for evaluation of bandwidth to be reserved to meet the QoS objectives. A new connection is accepted if the reserved bandwidth is less than the link capacity. The dynamic call admission control in [SS91] uses the measured number of cells arrived during a fixed interval and the declared traffic parameters to estimate the CLR.

Besaou et.al [BLCT97] proposed a fuzzy-based algorithm to predict the CLR. The fuzzy approximation estimates the cell loss ratio using: 1) the CLR when the system size (for example small buffer, low service rate etc.), 2) the asymptotic behavior of the CLR when the system size is large. The CAC mechanism (Figure 4.7) employs two components. The first one consists of a set of virtual buffers with reduced service capacity, to observe high cell loss with a small variance within a short measurement interval. The second component is a fuzzy approximation and a decision process. The fuzzy algorithm determines the required bandwidth for ongoing calls while the decision process makes a decision to accept or reject a new call based on the outputs of the algorithm.

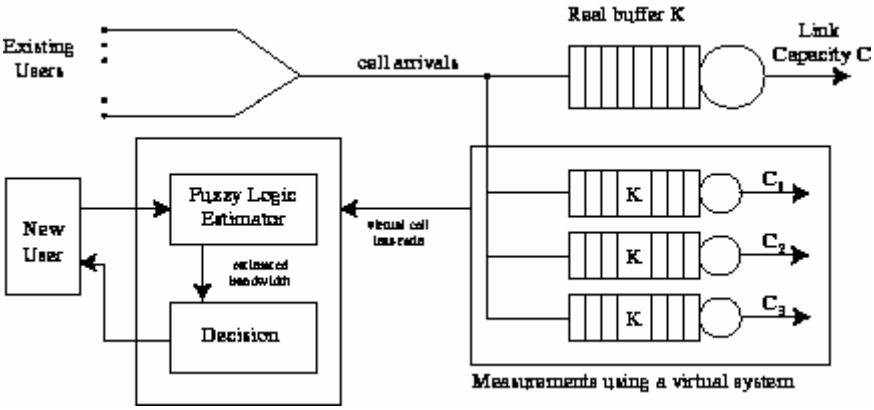

Figure 4.6. A CAC mechanism based on Fuzzy Logic

CAC FOR ABR AND UBR TRAFFIC

The connection admission control procedures for ABR and UBR are quite simple. The ABR traffic has a Minimum Cell Rate (MCR) guarantee. Thus, the CAC function has to assign this bandwidth for each connection. However, UBR traffic does not have such bandwidth

guarantees. Therefore, the CAC can admit many connections. Since UBR connections only get what's left over, the CAC function limits the number of UBR connections so that each connection gets some throughput in the long term. If the Minimum Cell Rate (MCR) guarantee for UBR (called UBR$^+$) is standardized, then the CAC function needs to allocate the bandwidth of MCR for each UBR connection as well.

TUNING THE CONNECTION ADMISSION CONTROL

Applications derive the required traffic descriptors based on some delay or response time requirements while trading off the cost of different connection bandwidth. It is quite unlikely that the connection will need the full amount of negotiated bandwidth on a continuous basis. For data applications, connections make use of the bandwidth for transferring large amounts of data. Smaller amounts of data can also be exchanged periodically, but the usage is generally much lower than the stated average rate. In the case of permanent connections (PVCs), the actual utilization compared to the negotiated sustained cell rate is often orders of magnitude smaller when measured over the lifetime of the connection. For switched connections (SVCs), this difference is usually smaller, since the connection is torn down when the transfer of information is completed.

The under-utilization of the allocated bandwidth coupled with the fact that the CAC allocates bandwidth conservatively can lead to severe loss of efficiency by the network. With appropriate measurement of the resource utilization, it is possible to quantify the amount of under-utilization and to tune the CAC function appropriately. CAC *booking or scaling factors* are generally available to allow the actual usage of the bandwidth to be taken into account, by artificially reducing the amount of bandwidth statically allocated for each connection, thus allowing more connections to be admitted. It should be noted that the statistical gain achieved through booking or scaling factors is completely different from the gains achieved through statistical multiplexing. In the former case, the gain is possible due to connections not utilizing their bandwidth, while in the latter, the gain is possible due to multiplexing many sources together.

There are two ways of tuning the CAC function, by applying some *over-booking factor* to the equivalent or virtual bandwidth calculated or by *scaling* (using *scaling factors*) the traffic descriptors before applying the calculations to derive the equivalent or virtual bandwidth. Since the relationship between the traffic descriptors and the allocated bandwidth is not linear, the use of the booking and scaling factors result in different over-booking levels. In both cases, if not carefully engineered, over-booking can result in loss of QoS guarantees for all the connections sharing the over-booked resource.

The scaling factor is engineered by measuring the usage of the bandwidth over a long period of time, on a per-connection basis. If it is determined that connections of a given range of traffic descriptors, or traffic coming from a given type of applications, do not use more than 50% of their SCR, then the SCR can be scaled down by at most 50%, leaving room for error and unpredictable trends. The scaled down value is only used for CAC purposes and is not used for policing, since the connection should be allowed to generate bursts

according to the negotiated traffic descriptors. If further measurements indicate that the usage is growing, or reducing, the scaling factors can be modified accordingly.

Booking factors are engineered by measuring the growth trends of the queues. Congestion measures are often available to provide insights as to whether the queue can be over-booked without affecting the QoS. Congestion measures indicate whether the queue size has exceeded specific thresholds for a given period of time. Based on that information, it is possible to decide on how aggressively the resource can be over-booked.

It is difficult to provide a definite recipe to perfectly tune the CAC in order to achieve very high utilization using only traffic with statically allocated bandwidth. The dynamics of an SVC-based network renders the task more complex, because measured statistics are likely to vary with the mix of connections at a given time. It is important to note that the under-utilization of the allocated resource can be at any time effectively used by bandwidth on demand services (ABR, GFR, UBR). Therefore the overall network can remain efficiently utilized.

REVIEW

The Connection Admission Control (CAC) is a very important function in ATM switches. The CAC function determines the admissibility of a new connection into the network by checking the availability of various system resources. The ATM network efficiency thus depends on how well the CAC function models the traffic and queuing behavior of the underlying congestion point. This chapter presented some of the approaches. As this area is most widely researched and published in the literature, it is difficult to enumerate or discuss all of them here. However, as a note of conclusion, a few other interesting approaches will be mentioned:

1. Gibbens et.al [GKK95] discuss admission procedures based on Bayesian rules, where acceptance decisions are based on whether the load is less than a pre-calculated load.
2. Duffield et.al [DLORT95] proposed to use an empirical entropy function to estimate QoS parameters, by-passing the modeling procedures.
3. A regression approach is used in [ROAG98], where simulation data is used to develop regression models for cell loss, delay and these estimates are used in computing the effective bandwidths.

References

[BC92] C.Blondia, O.Casals, “Statistical Multiplexing of VBR Sources: A matrix-analytical approach”, *Performance Evaluation* 16(1992), pp.5-20.

[BD94] E.Buffet, N.G.Duffield, “Exponential Upper Bounds via Martingales for multiplexers with Markovian Arrivals”, *Journal of Applied Probability*, 31, 1049-1061(1994)

[BLCT97] B.Bensaou, S.T.C.Lam, H.Chu, D.H.K.Tsang, “Estimation of the Cell loss ratio in ATM Networks with a Fuzzy system and Application to Measurement-based Call Admission Control”, *IEEE/ACM Transactions on Networking*, Vol.5, No.4, August 1997.

[CLW96] G.L.Choudhury, D.M.Lucantoni, W.Whitt, “Squeezing the Most Out of ATM”, *IEEE Transactions on Communications*, Vol. 44, No. 2, February 1996.

[CFW94] C.Courcoubetis, G.Fouskas , R.Weber, “On the Performance of an Effective Bandwidth Formula”, *Proceedings of the International Teletraffic Congress*, ITC14, 1994, pp.201-212

[CKW95] C.Courcoubetis, G.Kesidis, A.Ridder, J.Walrand, “Admission Control and Routing in ATM Networks using Inferences from Measured Buffer Occupancy”, *IEEE Transactions on Communications*, Vol.43, No.2/3/4, February/March/April 1995.

[Duf92] N.G.Duffield, “Rigorous Bounds for Loss Probabilities in Multiplexers of Discrete Heterogeneous Markovian Sources”

[DLORT95] N.G.Duffield, J.T.Lewis, N.O’Connell, R.Russell, F.Toomey, “Entropy of ATM Traffic Streams: A Tool for Estimating QoS Parameters”, *IEEE Journal on Selected Areas in Communications*, Vol.13, No.6, August 1995

[DJM97] Z.Dziong, M.Juda, L.G.Mason, “A Framework for Bandwidth Management in ATM Networks -–Aggregate Equivalent Bandwidth Estimation Approach”, *IEEE/ACM Transactions on Networking*, Vol.5, No.1, February 1997.

[DRS91] Lisa G.Dron, G. Ramamurthy, B.Sengupta, “Delay Analysis of Continuous Bit Rate Traffic Over an ATM Network”, *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 3, pp. 402-407, April 1991.

[EM93] A.Elwalid, D.Mitra, “Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks”, *IEEE/ACM Transactions on Networking*, Vol.1, No.3, June 1993.

[EMW95] A.Elwalid, D.Mitra, R.H.Wentworth, “A New Approach for Allocating Buffers and Bandwidth to Heterogeneous, Regulated Traffic in an ATM Node”, *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 6, pp. 1115-1127, August 95

[FLV94] G.Fiche, W.Lorcher, R.Veyland, F.Oger, “Study of Multiplexing for ATM traffic sources”, *International Teletraffic Congress ITC14*, pp. 441-452, June 1994.

[GAN91] R.Guerin, H.Ahmadi, M.Naghshineh, "Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 7, Sepetember 1991.

[GH91] R.J.Gibbens, P.J.Hunt, "Effective bandwidths for the multi-type UAS channel", *Queueing Systems 9 (1991)*, pp. 17-28.

[GKK95] R.J.Gibbens, F.P.Kelly, P.B.Key, "A Decision-Theoretic Approach to Call Admission Control in ATM Networks," *IEEE Journal on Selected Areas in Communications*, Vol.13, No.6, August 1995.

[Hui88] J.Y.Hui, "Resource Allocation for Broadband Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 6, No.9, pp.1598-1608.

[HW96] I.Hsu, J.Walrand, "Admission Control for Multi-Class ATM Traffic with Overflow Constraints," *Computer Networks and ISDN Systems Journal*, 28(13), pp.1739-1752, 1996.

[JSD97] S.Jamin, S.J.Shenker, P.B.Danzig, "Comparison of Measurement-based Admission Control Algorithms for Controlled-Load Service", *Proceedings of IEEE INFOCOM'97*, Kobe, Japan.

[Kel91] F.P.Kelly, "Effective Bandwidths at Multi-Class Queues", *Queuing Systems* 9 (1991), pp.5-16.

[KGC94] V.Kulkarni, L.Gun, P.Chimento, "Effective Bandwidth vector for Two-Priority ATM Traffic", *Proceedings of IEEE INFOCOM'94.*

[KWC93] G.Kesidis, J.Walrand, C.Chang, "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources", *IEEE/ACM Transactions on Networking*, Vol.1, No.4, pp.424-428.

[Nor94] I.Norros, "A Storage Model with self-similar Input", *Queueing Systems*, Vol.16, 1994, pp.387-396

[Nor96] I.Norros, "On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks", *IEEE Journal on Selected Areas in Communications*, Vol.13, No.6, August 1995.

[NRSV91] I.Norros, J.W.Roberts, A.Simonian, J.T.Virtamo, "The Superposition of Variable Bit Rate Sources in an ATM Multiplexer", *IEEE Journal on Selected Areas in Communications*, Vol.9, No.3, April 1991, pp.378-387.

[RK94] Q.Ren, H.Kobayashi, "Diffusion Process Approximations of a stastical multip[lxer with Markov Modulated Bursty Traffic Sources", *Proceedings of IEEE Globecom'94.*

[RMV96] J.Roberts, U.Mocci, J.Virtamo, "Broadband Network Teletraffic", Springer 1996.

[ROAG98] S.Ramaswamy, T.Ono-Tesfaye, W.W.Armstrong, P.Gburzynski, "Effective Bandwidths for Real-Time Traffic", To appear in *Journal of High-Speed Networks*, 1998.

[RV91] J.W.Roberts, J.T.Virtamo, "The Superposition of Periodic Cell Arrival Streams in an ATM Multiplexer", IEEE Transactions on Communications, Vol. 39, No. 2, pp. 298-303, February 1991.

[Skl94] A.Skliros, “A Connection Admission Control Algorithm for ATM Traffic Distorted by Cell Delay Variation”, *Proceedings of ITC14*, pp.1385-1394.

[SS91] H.Saito, K.Shiomoto, “Dynamic Call Admission Control in ATM Networks”, *IEEE Journal on Selected Areas in Communications*, Vol.9, No.7, September 1991.

[TG97] B.Tsybakov and N.D.Georganas, “Overflow Probability in an ATM queue with self-similar input traffic”, *Proceedings of IEEE ICC’97*, Montreal, June97.

[WK90] G.Woodruff, R.Kositpaiboon, “Multimedia Traffic Management Principles for Guaranteed ATM Network Performance”, IEEE Journal on Selected Areas in Communications, Vol. 8, No. 3, pp. 437-446, April 1990.

[YT95] T.Tang, D.H.K.Tsang, “A Novel Approach to Estimating the Cell Loss Probability in an ATM Multiplexer Loaded with Homogeneous On-Off Sources”, IEEE Transcations on Communications, Vol.43, No.1, January 1995, pp.117-126.

[ZT97] M.Zukerman, P.W.Tse, “An Adpative Connection Admission Control Scheme for ATM Networks”, IEEE conference ICC97, Montreal.