Blocking Probabilities in ATM Pipes Controlled by a Connection Acceptance Algorithm Based on Mean and Peak Bit Rates

E. Wallmeier, C. M. Hauber
Siemens AG, OEN ZL SV 23, Hofmannstr. 51, P.O.Box 70 00 73
W-8000 Munich 70, Germany

A connection acceptance algorithm for ATM pipes (i.e. trunks within an ATM network or links within an ATM exchange) is investigated which allows statistical multiplexing of heterogeneous traffic. It is demonstrated how to calculate the connection blocking probability of an ATM pipe controlled by this algorithm, using the concept of acceptance regions and blocking boundaries in conjunction with the generalized multidimensional Erlang loss formula. The performance of the algorithm is compared with that of peak bit rate reservation for several examples.

1. Introduction

In an Asynchronous Transfer Mode (ATM) network, the connection admission control (CAC) has to decide whether the idle bandwidth of a pipe (i.e. a trunk within the ATM network or a link within an ATM exchange) is sufficient to accept an incoming new connection without putting the Grade of Service (GoS) of any of the connections on the pipe outside specified limits. The decision to accept an additional connection depends on connection parameter values which have to be negotiated at connection set up time by the terminal equipment and the network. Throughout the holding time of the connection it has to be monitored that the source behaves within the agreed parameter values. The algorithm used by the CAC (i.e. the connection acceptance algorithm) should support statistical multiplexing in order to efficiently exploit the transmission capacity of an ATM pipe used by variable bit rate (VBR) connections. On the other hand, it has to be simple due to real time constraints and should use only such connection parameters which can be monitored by the network in an efficient way.

In the literature several connection acceptance algorithms have been discussed ([11] - [41]). The approach of [2] uses two parameters to characterize a connection: the mean bit rate and an upper bound for the bit rate variance defined with the aid of a "Gaussian envelope" of the source distribution. In [4] the mean bit rate, the peak bit rate and an upper bound for the bit rate variance calculated from these two parameters are used. In this paper we describe a generalized version of the connection acceptance algorithm presented in [4]. It is proposed to choose the upper bound for the bit rate variance dependent on the behaviour of the source. For a lot of video codecs the upper bound proposed in [2] may be adequate. For data traffic the upper bound calculated from the mean and peak bit rates seems more favourable. The algorithm presented here guarantees that the cell loss probability due to buffer overflow remains below a predefined value for every accepted connection. Waiting times are not considered here. However, the algorithm can be modified such that it is guaranteed that the waiting times exceed a critical value only with a given probability. In the second part of the paper it is shown how to calculate the connection blocking probability for the presented acceptance algorithm with the aid of the generalized multidimensional Erlang loss formula. The formula is applied to investigate the performance of the acceptance algorithm in the case of heterogeneous data traffic.

2. Characterization of ATM Connections

The connection acceptance algorithm considered here is based on a traffic model (see also [4]) that describes ATM connections on three levels which refer to different time scales: the cell level, the burst level and the connection level. On the cell level a connection is characterized by the distribution of the cell interarrival times. On the burst level a connection is described by the (instantaneous) bit rate which is considered as a function of time. Here it is assumed that the instantaneous bit rate at time t is given by a random variable X_t having a distribution which is independent of t. On the connection level a connection is characterized by the holding time and additional parameters, e.g. the expectation m=E(X) of the random variable X (mean bit rate of the connection) and the maximal value p which is assumed by X (peak bit rate of the connection). A bit rate, the instantaneous as well as the mean bit rate, can be measured counting the cells observed during a fixed length interval [5]. For measuring the mean bit rate, the length of the interval has to be longer than for measuring the instantaneous bit rate.

3. Cell Losses in an ATM Network

We investigate an ATM pipe which is connected as an outlet to an ATM multiplexer or switching element. If more than one cell arrives at the same time for the considered outlet the cells which cannot be transmitted have to be buffered instead. In the case of buffer overflow this leads to cell losses. The instantaneous cell loss probability (i.e. the probability of loosing an arriving cell) because of buffer overflow at time t results from two types of congestion:

1) cell level congestion which is not evident on the burst level;
2) burst level congestion.
We speak of cell level congestion, when a cell arriving at time t is lost although the sum of the instantaneous bit rates produced by the sources does not exceed the "permissible load" which is defined as follows: Let \( v_i(x_1, ..., x_n) \) denote the probability that a cell of connection i (where \( 1 \leq i \leq n \)) arriving at time t is lost owing to buffer overflow, given that n connections send with instantaneous bit rates \( x_1, ..., x_n \) at time t. It is assumed that this conditional probability is independent of i. This assumption is justified if the different inlets of the considered switching element are served using a fair strategy. The subscript i is omitted in the following. For a given value \( Q^* \) (denoting the permissible cell loss probability due to cell level congestion), the corresponding permissible load \( C(Q^*) \) is defined as the largest value (measured in Mbit/s) satisfying the following condition: For an arbitrary number n of arbitrary connections which may use the pipe and every t

\[
v_i(x_1, ..., x_n) \leq Q^* \text{ if } \sum_{i=1}^{n} x_i \leq C(Q^*). \tag{1}
\]

Cell level congestion is mainly a consequence of the variable cell delay produced by the ATM network itself. Variable cell delay arises due to the asynchronous multiplexing and the set up and release of connections even if only cells of constant bit rate connections (CBR) are to be transmitted.

**Burst level congestion** occurs if the sum of the instantaneous bit rates of the sources exceeds the permissible load.

The permissible load \( C(Q^*) \) can be determined calculating \( v_i(x_1, ..., x_n) \) and checking inequality (1) while increasing the load \( \sum_i x_i \) step by step. In the literature a lot of cell level models have been proposed which allow to calculate the cell loss probability \( v_i(x_1, ..., x_n) \) for an ATM pipe carrying a fixed number of CBR connections [6], [7], [8]).

If a new connection is accepted only if the sum of the peak bit rates of all n connections including the new one is less than or equal to \( C(Q^*) \), i.e. if

\[
\sum_{i=1}^{n} p_i \leq C(Q^*), \tag{2}
\]

then the new connection is accepted. Otherwise the incoming new connection is accepted if both

\[
c = C(Q^*) - \sum_{i=1}^{n} p_i \geq \bar{c}_{II}, \tag{3}
\]

i.e. the bandwidth c available for class II connections allows statistical multiplexing, and

\[
q(c) \cdot \left( \sum_i \sigma_i^2 \right)^{1/2} \leq \sum_{i=1}^{n} m_i + \max_{i=1}^{n} p_i \leq C(Q^*) \tag{4}
\]

i.e. the bandwidth required for class II connections is less than or equal to c.

The values \( \bar{c}_{II} \) and \( q(c) \), \( c \in [C_{II}, C(Q^*)] \), are parameters of this acceptance algorithm which have to be determined in advance (with the aid of a computer program). They depend on the permissible cell loss probability \( Q_{max} \), the \( Q^* \) chosen and the definition of the two classes of connections. In section 4.3 conditions for these parameters are presented under which the algorithm guarantees that the instantaneous cell loss probability of every connection due to buffer overflow remains below the value \( Q_{max} \) for an arbitrary t.
4.2 Upper Bounds for the Variance

The algorithm presented here uses three connection parameters which have to be negotiated at call set up time and which have to be policing throughout the holding time of the connection: the peak bit rate p, the mean bit rate m and an upper bound $\sigma^2$ for the variance of the distribution of the instantaneous bit rate $X_t$. In order to allow a practical implementation of the algorithm, it is necessary to choose the upper bound for the variance in a way that it is possible for the terminal equipment to specify a value at call set up time and that it is possible for the network to police this value in an efficient way. The statistical multiplexing gain is influenced by the difference between the bit rate variance of the source and the upper bound which is used for practical purposes. Two different upper bounds are discussed below. Which of these upper bounds is the most efficient depends on the source behaviour. It may be favourable to use different upper bounds for different services.

An Upper Bound Calculated from the Mean Bit Rate and the Peak Bit Rate

The variance of a connection with bit rate $X_t$ mean bit rate $m$ and peak bit rate $p$ satisfies the following inequality:

$$\text{Var}(X_t) = E(X_t^2) - (E(X_t))^2 \leq E(p - X_t)^2 - (E(p - X_t))^2 = m^2(p - m). \quad (5)$$

The upper bound $m^2(p - m)$ coincides with the bit rate variance if and only if the connection has on/off characteristic, i.e. has only two states: sending at peak bit rate or not sending at all. Using this upper bound, only the two parameters mean bit rate $m$ and peak bit rate $p$ are needed to characterize a connection. These two parameters can be monitored by a dual "Leaky Bucket" if the sources use a traffic shaper to control their mean bit rates [9]. Without using traffic shapers it is practically not possible to police the mean bit rate without unjustified cell discarding [10]). In particular for data connections the use of this upper bound seems favourable.

An Upper Bound for Approximately Gaussian Source Distributions

For many video codecs the bit rate variance may be badly overestimated if the upper bound calculated above is used. [11] contains an example of a video codec where no statistical multiplexing gain can be achieved if this upper bound is used for a 150 Mbit/s ATM pipe. If the bit rate of the source is approximately Gaussian distributed (for video codecs it may be possible to achieve such a distribution with the aid of an appropriate traffic shaper), then the mean bit rate and the variance of a "Gaussian envelope" of the negative cumulative distribution function $P(X > x)$ can be used instead of the exact values; i.e. the mean bit rate and the variance of a Gaussian negative cumulative distribution function $F(x)$ satisfying $F(x) > P(X > x)$ for all $x \geq E(X)$ (see [2]). In [5] a policing method called "Gabart policing" is described which allows to monitor whether the behaviour of the source fits to the Gaussian envelope. If the bit rate of a traffic source is not approximately Gaussian distributed, the difference between the source distribution and the Gaussian envelope (which can be policed more easily) may become so large that it is more favourable to use the upper bound calculated above.

4.3 Calculation of the Parameters of the Algorithm

For the computation of the parameters $C_i$ and $q(c)$, $c \in [C_m, C(Q^*)]$, the concept of a "critical region" is introduced:

Any set of connections is defined as belonging to the critical region if the following applies:

1) The connections are accepted.
2) No connection is accepted in addition to these.
3) The sum of their peak bit rates exceeds $C(Q_{max})$.

By slightly modifying the proofs used in [4] it can be shown that the instantaneous cell loss probability $Q_i$ of a connection $i$ at time $t$ using a pipe occupied by $n$ connections (with peak bit rates $p_i$) satisfies the following inequality:

$$q(c) = \frac{E(X_{t,i})}{n} \leq \frac{C(Q^*) - p_i}{C(Q^*) - m} \quad (6)$$

Further it can be shown in the same way as in [4] that the right hand side of (6) and thus $Q_i$ is less than or equal to the permissible probability $Q_{max}$ for every $i$ and $t$ if the bit rates of all connections are uncorrelated and if the values $q(c)$, $c \in [C_m, C(Q^*)]$ are chosen so large that for every set of $n$ connections belonging to the critical region the following condition is fulfilled (assume that the connections have peak bit rates $p_i$, mean bit rates $m_i$ and that connections 1,...,r belong to class II and connections $r+1,...,n$ to class I):

For

$$c = C(Q^*) - \sum_{i=1}^{r} m_i, \quad r = \frac{\sum_{i=1}^{r} \text{Var}(X_{t,i})}{\sum_{i=1}^{n} \text{Var}(X_{t,i})}$$

the inequality

$$\sum_{i=1}^{r} \frac{\text{PC}(\sum_{i=1}^{r} X_{t,i} - m) / a > q(c)) \leq Q_{max} - q(a) \quad (7)$$

applies.

According to the central limit theorem, the distribution of the bit rate of the superposed traffic obeys approximately a Gaussian distribution if the number of multiplexed sources is large. If, for every set of connections which belongs to the critical region, the aggregate bit rate of the class II connections had a Gaussian distribution, then the $q(c)$ could be chosen as an $\alpha$-quantile of the standardized normal distribution for $\alpha = Q_{max} - Q^*$ (e.g. $q(c) \approx 6.4$ for $Q_{max} = 10^{-10}, Q^* = 10^{-11}$) and would be independent of $a$. In this case the equal sign holds in inequality (7).

Generally however the assumption of a Gaussian distribution is not justified. Therefore a "worst case" distribution within the critical region has to be determined. For given classes I and II, values $q(c), c \in [C_m, C(Q^*)]$ have to be postulated and then inequality (7) must be checked. In this case the $\alpha$-quantile of the standard normal distribution serves as a lower bound for finding values $q(c)$.

In the following examples we make the assumption that $Q_{max} = 10^{-10}$. We choose $Q^* = 10^{-11}$ and assume that $C(Q^*) = 0.85 \cdot (45/53) - 149,760$ Mbit/s...
5.1 Acceptance Regions and Blocking Boundaries

To write down the generalized multidimensional Erlang loss formula, it is necessary to introduce the concept of acceptance regions and blocking boundaries [12], [7].

A connection acceptance algorithm defines connection types. Two connections are said to belong to the same connection type if the connection parameters (e.g., mean bit rate and peak bit rate) which are forwarded to the acceptance algorithm at call set up time have the same values.

Consider an ATM pipe which is offered a traffic mix of k connection types. The acceptance region S is defined as a set of k-tuples \( n=(n_1,n_2,...,n_k) \) (called occupancy states) satisfying the following condition:

\[ S = \{ n | \text{the connection acceptance algorithm} \}
\]

In other words: Given the occupancy state \( (n_1,n_2,...,n_k) \), the connection acceptance algorithm accepts an incoming new type \( j \) connection if and only if \( (n_1,n_2,...,n_j+1,...,n_k) \in S \). If \( (n_1,n_2,...,n_j+1,...,n_k) \not\in S \), then the occupancy state \( (n_1,n_2,...,n_j+1,...,n_k) \) is called a blocking state for type \( j \) connections. The blocking boundary \( T_j \) for type \( j \) connections is defined as the set of all blocking states for type \( j \) connections:

\[ T_j = \{ (n_1,n_2,...,n_k) | \text{for } j \} \]

Let us first consider the acceptance region and the blocking boundary of an ATM pipe carrying homogeneous traffic, i.e. connections of only a single connection type \( (k=1) \). In this case there is a \( n_{1\text{max}} \) such that

\[ S = \{ n_j | 0 \leq n_j \leq n_{1\text{max}} \} \]

\[ T_1 = \{ n_{1\text{max}} \} \]

For a connection type having a peak bit rate of \( p_j = 2.048 \text{ Mbit/s} \) and a mean-to-peak bit rate ratio of \( m_j/p_j = 0.1, 0.2, ..., 0.9, 1.0 \), we get the following values of \( n_{1\text{max}} \) by applying the Sigma Rule, case a):

<table>
<thead>
<tr>
<th>( m_j/p_j )</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_{1\text{max}} )</td>
<td>199</td>
<td>105</td>
<td>74</td>
<td>59</td>
<td>56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

That means that in this example a statistical multiplexing gain can be achieved only for a mean-to-peak bit rate ratio below 0.5.

Figure 1 shows acceptance regions for further examples. Acceptance regions are the sets of all pairs \( (n_1,n_2) \) under the respective curves.

![Figure 1: Acceptance regions (n_1, n_2).](image-url)
5.2 The Generalized Multidimensional Erlang Loss Formula

Consider an ATM pipe carrying a mix of k connection types. Using the above introduced notation of an acceptance region S and blocking boundaries T, the blocking probability B of type j connections (i.e. the probability that an incoming new type j connection has to be rejected) can be calculated with the aid of the generalized multidimensional Erlang loss formula

\[
B_j = \frac{\sum_{n=1}^{\infty} A_1^{n_1} A_2^{n_2} \cdots A_k^{n_k}}{\sum_{n=1}^{\infty} \prod_{j=1}^{k} (A_j + n_j^{\infty})} \quad j = 1,2,\ldots,k \tag{10}
\]

where \( A_j \) denotes the offered traffic of type j connections (measured in Erlang).

This formula is valid if the arrival process of incoming new type j connections is a stationary Poisson process with arrival rate \( A_j \) and if the distribution of the holding time of type j connections has a rational Laplace transform \([12],[7]\).

The offered traffic of type j connections then amounts to \( A_j = \frac{1}{\mu_j} \). We furthermore define for later use the offered traffic volume of type j connections as \( V_j = A_j \mu_j \) and the total offered traffic volume \( V = V_1 + V_2 + \cdots + V_k \) (both measured in Erlang/s).

Let us again investigate the blocking probability of an ATM pipe that carries homogeneous traffic \( k = 1 \). We compare the following examples: Connections of the single connection type on the ATM pipe have either a peak bit rate of \( p_1 = 2.048 \text{ Mbit/s} \), a peak bit rate of \( p_2 = 4.096 \text{ Mbit/s} \) or a peak bit rate of \( p_3 = 6.144 \text{ Mbit/s} \). They have a mean-to-peak bit rate ratio of \( m_j/p_j = 0.1 \). As connection acceptance algorithm we apply either one of the variants of the Sigma Rule (i.e. case a), b) or c)) or peak bit rate reservation.

Figure 2 depicts for these examples the connection blocking probability \( B_i \) of the single connection type on the ATM pipe as a function of the total offered traffic volume \( V \). The solid curves show the results for the three variants of the Sigma Rule (cases a), b), c)), the dashed curves those for peak bit rate reservation (PBRR).

We next investigate the blocking probability of a single connection type on the ATM pipe that carries a special mix of heterogeneous traffic, namely one of the mixes of two types of connections already taken as a basis for figure 1: \( p_1 = 64 \text{ kbit/s}, p_2 = 2.048 \text{ Mbit/s}, m_1/p_2 = 0.1 \). We compare five different examples which we get for
a type 2 offered traffic volume in proportion to the total offered traffic volume of $V_2/V = 90\%, 70\%, 50\%, 30\%, 10\%$ (and correspondingly $V_1/V = 1 - V_2/V$). As connection acceptance algorithm either the Sigma Rule, case a), or peak bit rate reservation is applied.

Calculation of blocking probabilities using the generalized Erlang loss formula (10) generally requires a substantial numerical effort. However, for the two-dimensional case considered here it was possible to calculate the acceptance region $S$, the blocking boundaries $T_1$, and $T_2$ and the blocking probabilities $B_1$ and $B_2$ on a personal computer.

Figure 3 depicts the blocking probability $B_2$ of type 2 connections as a function of the total offered traffic volume $V = V_1 + V_2$ for the five different examples mentioned above. The solid curves show the results for the Sigma Rule, case a), the dashed curves those for peak bit rate reservation (PBRR).

![Figure 3: Connection blocking probability $B_2$ of the second connection type in the case of heterogeneous traffic as a function of the total offered traffic volume $V$.](image)

In this special example the blocking probability $B_2$ (not displayed) is at any point smaller than the blocking probability $B_1$. This is due to the fact that in this example the blocking boundary $T_2$ for type 1 connections is a subset of the blocking boundary $T_2$ for type 2 connections.

With peak bit rate reservation - dashed curves of figure 3 - we observe for moderate values of the total offered traffic volume ($V = 60 \ldots 100$ ErlMbit/s) that the smaller the proportion of type 2 traffic becomes, the steeper is the slope of the blocking curves and the smaller is the blocking probability. With increasing total offered traffic volume all curves finally approach the value $B_2 = 1$ asymptotically which is a natural property of blocking curves of loss systems.

With the Sigma Rule - solid curves of figure 3 - the traffic handling capability increases with growing proportion of type 2 VBR traffic which can be statistically multiplexed in contrast to type 1 CBR traffic.

**Conclusion**

In this contribution we described a connection acceptance algorithm which shows that statistical multiplexing of heterogeneous ATM traffic is feasible under the condition that the traffic sources shape their output according to connection parameters which have been negotiated with the network. The results confirm that the statistical multiplexing gain is considerable if the peak bit rate of most of the connections is small compared with the total net transmission bit rate of the pipe and if the peak-to-mean bit rate ratio (burstiness) of these connections is high. A modified version of the Sigma Rule, making use of explicit cell loss priorities, is currently under study.

**Acknowledgement:**

This work has been supported by the Commission of the European Communities under RACE contract R1012 "Broadband Local Network Technology".

**References:**