SIMPLE DECISION RULES FOR ACCEPTANCE OF MIXED TRAFFIC STREAMS.

Frits C. SCHOUTE

Philips Telecommunication and Data Systems
Hilversum, The Netherlands.

Mixed traffic means different bandwidth requirements for different calls. With Asynchronous Time Division (ATD), the calls will use virtual connections. In this context it is not obvious whether a new call can be accepted or not.

In this paper the problem is analysed by first introducing a three-level characterization of a traffic stream. Next we study how much loading of an ATD channel is permitted at the TRANSMISSION level. This leads finally to a simple rule for acceptance of calls at CONNECTION level. This rule takes also into account the statistical character of the call at the intermediate ACTION level.

1. INTRODUCTION.

In the days that each call was a voice call which used a connection for some holding time, the decision to accept a new call, or not, was obvious. On a trunk of N lines, one accepts a new call only if the number of occupied lines is less than N. Erlang made a significant contribution by giving a formula to compute the blocking probability for a new call.

As the telephony network now grows out to carry not only voice communication, but also many forms of data communication via the Integrated Services Digital Network (ISDN), some new aspects are introduced. In the first place: different calls need different bandwidths (bits per second). Secondly, the maximum bandwidth used by a call, need not be allocated for the whole duration of a call. This is particularly true if ATD is used for transmission and switching.

An ATD channel can be viewed as a single server used by many traffic streams, where each stream generates some flow of cells (fixed length packets). The mixed traffic streams share the ATD channel on a statistical basis. Since the channel will typically not be loaded to a full 100%, it seems there is always room for one more call that uses only a small fraction of the channel capacity. However, at some point the loading will be so high that accepting another call will cause excessive waiting times for ATD cells.

To analyse the question of call acceptance, we must first give a statistical characterization of traffic streams. It is important to realize that different time scales are involved. E.g. a connection can last for more than 1000 seconds and a cell transmission takes less than 0.1 second. Therefore we introduce in section 2 a three-level characterization of traffic streams.

The permitted loading of an ATD channel depends on the delay requirements and the statistical properties of the arrival and service processes. In section 3 we analyse this relation.

In the last section before the conclusions, we show how the three-level description is useful in constructing a simple decision rule for acceptance of calls.
2. CHARACTERIZATION OF TRAFFIC STREAMS.

A call can be, for example, a voice call, a batch-data call or an interactive-data call. Without silence detection, the voice call is active for the whole duration of the connection. Considering facsimile transmission as an example of a batch-data call, we have that the call becomes active for each sheet (of one document) to be transmitted. With interactive-data, the call has short bursts of activity for each query/response pair.

We consider here the case that those traffic streams are carried by an ATD channel, which transmits fixed sized cells at some high bit-rate. Each call will generate cells during its active periods, at a rate corresponding to the desired bit-rate.

In the examples mentioned, one can recognize three levels of description. A similar distinction into three aspects of a traffic stream was made by Ven [1]. Each level has its own time scale. The slowest level is the (virtual) CONNECTION level. At an intermediate time scale we have the ACTION level. For an ordinary voice call, the active period lasts as long as the connection. If silence detection is used, the active period would last for the duration of a talk spurt. At the lowest level, the TRANSMISSION level, we see for each call an on/off sequence. The on-time lasts exactly the duration of one packet transmission.

Figure 1 illustrates the duration of the connection, the active periods and the cell transmissions of a call. The active periods, take only place when there is a connection (i.e. the line at CONNECTION level is high). In the same way, the on/off sequence at TRANSMISSION level is only present during the active period of a call.

![Figure 1. Three levels of characterization of a traffic stream.](image)

Our method of characterization of a traffic stream has several applications:

a. Formulation. One can see that the queueing delays for the ATD channel take place at the TRANSMISSION level. However the acceptance or blocking of calls takes place at CONNECTION level.

b. Simulation. Having, in a discrete event simulation, all three time scales in one model, leads to poor convergence of confidence intervals.

c. Computation. Here, three time scales in one model leads to numerical instability. It is better to do computations at one time scale, assuming things constant at the slower time scale. Events of the faster time scale can be summarized as probabilities.

4.2A.5.2
d. Specification. To specify a traffic stream one gives for each level the duration of the on-time and the duration of the off-time. Generally, these times will be stochastic variables so also the probability distribution will be given.

We continue with the aspect of specification of a traffic stream. We assume that, besides the on/off times, for each traffic stream there is a maximum delay $D$. Only with a small probability $\varepsilon$, also to be specified, this delay may be exceeded.

Let $S$ be the channel transmission speed in bits per second and $B$ the fixed cell size in bits. These two parameters determine $T_{on}$, the on-time at TRANSMISSION level. Alternatively, one can say that whenever two of the three parameters are given, the other one will follow from:

$$T_{on} = \frac{B}{S}$$ (2.1)

The specification of a traffic stream, will be a list containing the following parameters:

- $C_{on}$: On-time at CONNECTION level
- $C_{off}$: Off-time at CONNECTION level
- $A_{on}$: On-time at ACTION level
- $A_{off}$: Off-time at ACTION level
- $T_{on}$: On-time at TRANSMISSION level
- $T_{off}$: Off-time at TRANSMISSION level
- $D$: Maximum delay at TRANSMISSION level
- $\varepsilon$: Probability that cell delay exceeds $D$
- $B$: Cell size in bits

The on/off times are mean values of some probability distribution which may also be specified.

3. PERMITTED LOADING OF AN ATD CHANNEL.

In this section we consider the traffic streams at the TRANSMISSION level. At this level cell transmission takes place. A cell is a fixed length packet of e.g. 2+32 octets. The first two octets form the header, indicating the virtual circuit to which the cell belongs. This fixed cell size was proposed at the Hamburg meeting of CCITT committee XVIII [2].

Simulation runs of an ATD channel showed that up to a high level of channel loading (90%), no measurable delay excess occurs. The deterministic arrival statistics of the voice cells and, to a lesser degree, of the data-batch cells mainly accounts for this behaviour.

This leads to a more fundamental study of some standard queueing systems to determine the relation between permitted loading and the statistics of the arrival and service processes. We use the Kendall notation to name the queueing systems in short-hand [3]. We use the letter $U$ to denote the uniform distribution.

The five systems we studied can be arranged in the following diagram.

<table>
<thead>
<tr>
<th>Coefficient of variation of interarrival time:</th>
<th>Coefficient of variation of service time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$D/M/1$</td>
<td>$M/M/1$</td>
</tr>
<tr>
<td>$D/D/1$</td>
<td>$U/D/1$</td>
</tr>
<tr>
<td></td>
<td>$M/D/1$</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

4.2A.5.3
The M/M/1 queue is included because it is the simplest and most frequently used model. But it is not the most applicable model for our ATD channel, because we have a service time that is deterministic and the interarrival time is less random than negative exponential. The models D/M/1 and M/D/1 show what happens if either the arrival or the service process is made deterministic. D/D/1 is the extreme case, without random fluctuations. Maybe the most appropriate model is the U/D/1. Here it is modelled that the interarrival times are not deterministic, nevertheless there is some maximum on the interarrival time. The coefficient of variation of the uniform distribution is 1/\sqrt{3}.

To compare the five models by permitted loading, we must set some criterion. The criterion we use is:

\[ \Pr(\frac{T_q}{T_s} \leq 10) = 0.99 \]  

The sojourn time \( T_q \) is the sum of the wait time \( T_w \) and the (mean) service time \( T_s \). The criterion expresses that the sojourn time should exceed 10 times the mean service time only with 1% probability.

Space does not permit us to give the derivations of maximum permitted loading with the above criterion. In some cases the derivation is straightforward, in other cases it is more elaborate. For the U/D/1 queue the result was not obtained analytically, but by simulation. The results are given below.

<table>
<thead>
<tr>
<th>queue</th>
<th>permitted loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/M/1</td>
<td>0.539</td>
</tr>
<tr>
<td>D/M/1</td>
<td>0.746</td>
</tr>
<tr>
<td>M/D/1</td>
<td>0.773</td>
</tr>
<tr>
<td>U/D/1</td>
<td>0.908</td>
</tr>
<tr>
<td>D/D/1</td>
<td>0.999</td>
</tr>
</tbody>
</table>

**TABLE 1.** Permitted loading for different queueing systems. The U/D/1 queue is most representative for an ATD channel.

4. DECISION RULE FOR ACCEPTANCE OF CALL.

In this section we shall analyse decision rules at the CONNECTION level. The decision rule will take into account the behaviour of the traffic streams at ACTION and TRANSMISSION level.

From the previous section we know that the channel can carry 90% loading at TRANSMISSION level. However, the decision to accept a new call is made at CONNECTION level. In between we have the ACTION level, where the call can not be controlled. It is the uncertainty about the number of active virtual connections which makes a careful decision rule necessary. This decision rule will keep the average loading of the ATD channel considerably lower than the above mentioned 90%.

Van Arem [4] has also derived a decision rule, based on a more detailed model of the channel at TRANSMISSION level. In his model there is a frame structure, where each traffic stream gets a certain number of slots per frame. This decision rule has been compared with the one below. The comparison was based on how many of each type could be admitted. Because in [4] each traffic stream has its own slots, that decision rule is slightly more restrictive.
4.1. Decision rule based on m traffic streams

The loading of the channel is determined by the number of active virtual connections. The decision rule that we propose is such that it keeps the probability of too many active virtual connections below the given threshold $\epsilon$. We can write this as:

$$\Pr\{ \text{loading} > 90\% \} < \epsilon$$  \hspace{1cm} (4.1)

This probability depends of course on the number of virtual connections and the fraction of time that they are active. Therefore we introduce the following parameters:

- $N_i$: Number of virtual connections of type $i$, $i=1,2,\ldots,m$.
- $x_i$: Number of active virtual connections.
- $r_i$: Fraction of time that connection is active.
- $\rho_i$: Fraction of channel capacity used by one virtual call, when active.

One can directly express the parameters $r_i$ and $\rho_i$ in the traffic characterization parameters (2.2).

Given the numbers of active virtual connections, $x_i$, the loading of the channel is equal to:

$$\sum_{i=1}^{m} x_i \rho_i$$  \hspace{1cm} (4.3)

The activity of a virtual connection at an arbitrary point in time, is a binary random variable. We take the active periods of the different virtual connections as being statistically independent. The probability distribution of the number of active virtual connections is then a product of binomial distributions:

$$p(x) = \prod_{i=1}^{m} \binom{N_i}{x_i} r_i^{x_i} (1-r_i)^{N_i-x_i}$$  \hspace{1cm} (4.4)

where $x$ denotes the vector $[x_1,x_2,\ldots,x_m]$.

A subset of the space of possible values for $x$ is called the overload set $S_{.90}$. It is defined by:

$$S_{.90} = \{ x \mid \sum_{i=1}^{m} x_i \rho_i > 0.90 \}$$  \hspace{1cm} (4.5)

We want to maintain delay excess acceptable by keeping loading below 0.90 with probability $\epsilon$. The decision rule that accomplishes this is as follows.

Accept a new connection of type $i$ if, after $N_i := N_i + 1$ and recalculation of (4.4), the following inequality is still valid:

$$\sum_{S_{.90}} p(x) < \epsilon$$  \hspace{1cm} (4.6)

It will be clear that this decision rule will keep the average loading at a level well below 90%. We shall now investigate what a typical level of average loading is. This can then lead to an even simpler decision rule that just checks whether the average loading is below a certain threshold.
4.2. Decision rule based on two traffic streams

We introduce the following simplifications. Let \( N \) be the total number of virtual connections. All connections are assumed to have the same value of \( p \) (fraction of channel capacity used). We fix the value of \( r_i \) to 0.5. This is a worst case assumption, because it maximizes the variance of the binary random variable that indicates whether a connection is active. Later the worst case assumption will be weakened, by treating streams with \( r_i \) close to 1 separately.

The loading of the channel as determined by the number of active virtual connections is a random variable. With the above simplifications the expected value of channel loading (average loading), \( R \), is:

\[
R = \frac{N p}{2}
\]  
(4.7)

We shall determine below, the permitted value of \( R \). A binary random variable that can assume values 0 and \( p \), has 0.25 \( p^2 \) as maximum variance. From this it follows that the variance of channel loading, \( \sigma^2 \), is:

\[
\sigma^2 = \frac{R^2}{N}
\]  
(4.8)

By approximating channel loading as a Gaussian random variable \( X \), we can write for the probability that \( X \leq 0.90 \):

\[
\Pr\{ X \leq 0.90 \} = \text{Erf}(\frac{0.90-R}{R}) \]  
(4.9)

We still have the requirement that loading excess may occur only with a small probability \( \epsilon \). Hence:

\[
\text{Erf}(\frac{0.90-R}{R}) = 1 - \epsilon
\]  
(4.10)

Define:

\[
Q = \frac{0.90-R}{R} \]  
(4.11)

Then we can write, using Gitnik's approximation of the error function [5]:

\[
1 - e^{-kQ^2} = 1 - 4\epsilon - 4\epsilon^2
\]  
(4.12)

Here the constant \( k = 0.619536 \ldots \). Dropping the term of order \( \epsilon^2 \), substituting the right hand side of (4.11), and doing some algebra leads to:

\[
R = \frac{0.90}{1 + \left( -\ln(4\epsilon) \right)^{0.5} / k N}
\]  
(4.13)

For convenience we rewrite (4.13) using the base 10 logarithm and substituting k's value:

\[
R = \frac{0.90}{1 + \left( -\log(\epsilon) - 0.60 \right)^{0.5} / 0.27 N}
\]  
(4.14)
Relation (4.14) gives a simple expression for the permitted average loading on an ATD channel. One can see that as \( N \), the number of virtual connections becomes larger then the permitted value \( R \) comes closer to 0.90. The same is true when \( c \) is made larger.

The above results are based on the assumption that at a call is maximally unpredictable at ACTION level. For some calls this far too pessimistic, particularly for those calls that are active for the whole duration of the call.

Therefore, let us make distinction between two types of calls: type 1 is active for the duration of a call \( (r_1=1) \), and type 2 has \( r_2 < 1 \). The decision rule is then as follows:

Determine \( R \) from (4.14), using for \( N \) the number of type 2 connections. Let \( R_i \) be the average capacity that will be assigned to type \( i \) connections \( (i=1,2) \). Then that assignment will be acceptable if:

\[
R_2 \leq R \left( 1 - \frac{R_1}{0.90} \right)
\]  

(4.15)

5. CONCLUSIONS

For a proper description of mixed traffic streams, it is essential to distinguish different time scales. This leads to a three-level description, involving, the CONNECTION, ACTION and TRANSMISSION level. Queueing delays occur at TRANSMISSION level, acceptance decisions take place at CONNECTION level.

Simulation runs of an ATD channel and analysis of systems showed that at TRANSMISSION level 90% loading can be permitted. In the last section two simple decision rules translate the requirement at TRANSMISSION level into acceptance / blocking decision at CONNECTION level. The second decision rule requires even less computation than the first one.

REFERENCES


