CELL-LEVEL STATISTICAL MULTIPLEXING IN ATM NETWORKS: ANALYSIS, DIMENSIONING AND CALL-ACCEPTANCE CONTROL w.r.t. QOS CRITERIA

N.M. Mitrou and D.E. Pendarakis†

Abstract

A cell-level statistical multiplexing scheme supporting two service classes with different delay and cell-loss QOS requirements is investigated. Each service class is assigned certain priority levels on using the multiplexer resources (outlink bandwidth and buffer space) which monitor the quality of the received service. A three-fold problem is solved: The analysis, which consists in finding the delay probability distribution and the cell-loss probability for each service class in terms of the multiplexer dimensions, the load and the priority levels; the resource dimensioning and allocation problem, which is formulated as "given the desired operation region on the load plane and the QOS values for both classes, find the minimal required capacity of the multiplexer resources as well as the allocation parameters (or priority levels)"; finally, the Call-Acceptance-Control (CAC) problem, which is answered by specifying appropriate region on the load plane (a superset of the operation region) within which the QOS requirements could be satisfied through suitable allocation of the resources. Two approaches are followed: An approximate one, which leads to closed-form solutions and an exact one by means of a 2-D Markov model which is solved through numerical manipulations.

1 Introduction

The salient features of an Asynchronous Transfer Mode (ATM) network, currently being studied within many Research and Development projects as well as by standardization bodies, are the integration of services with different traffic profiles and different Quality of Service - QOS - requirements and the statistical type of multiplexing of information streams down to the packet (cell) level [1]. One of the most difficult problems which arise in such a network is how to dimension and allocate the network resources (outlink bandwidth and buffer space at every multiplexing & switching node) to the various services, in order to guarantee the required QOS with a minimal cost.

The well-known complete partitioning or complete sharing, non-prioritized schemes seem to be inefficient to satisfy both of the above conflicting demands [3], [4]. Therefore, other allocation mechanisms, which could utilize some kind of priorities, have to be devised and efficiently implemented. Moreover, a hierarchical functioning of such mechanisms seems to be necessary in order to cope with the multilevel resource allocation required by an ATM network, viz. from the connection level down to the cell level [5],[6].

In this paper a generic time & space resource allocation scheme at a cell level is modelled and analysed, with two service classes (CCITT is recommending one bit in the cell header for priorities,[2]). With an approximate modelling, simple formulas are derived giving the delay and the buffer-occupancy distributions. Using this simple toolset, the reverse problem of resource dimensioning and allocation is also solved in a closed form. A second approach is also followed through establishing a 2-Dimensional Markov model and solving it numerically, which yields the accurate solution.

2 Problem Formulation and Analysis

The statistical multiplexer we consider, is furnished with a time resource (time slots) and a space resource (buffer space) shared by two service classes. The following resource-sharing scheme is investigated:

- **Time resource allocation**: As far as there are pending requests (waiting cells) from both service classes, the resource (time slot) is allocated in a probabilistic way, i.e. with a probability $p_1 = p$ to service 1, otherwise (with $p_2 = 1 - p$) to service 2. If one service is idle, the slot is given to the other.

- **Space resource allocation**: The total capacity $L$ (in number of cells) of the space resource is divided into portions $l_1$ and $l_2$, which are assigned to service classes 1 and 2 respectively. As far as the capacity $L$ has not been exhausted, there are no cell losses. If, on the other hand, the buffer is full, an arriving cell is lost whenever the service class it belongs to has exceeded its limit $l_i$ ($i=1$ or 2).

---

*This work was partially funded by the EEC RACE 1022 project (Technology for ATD)

†National Technical University of Athens, Department of Electrical Engineering, Division of Computer Science, 157 73 Zographou, Athens, Greece, tel +3017759388, fax +3017757501 (manual).
Obviously, the resource allocation parameters $p_i$ and $\frac{1}{p_i}$ constitute the generalized time & space priority levels for the service class $i$. Figure 1 gives a schematic representation of the resource allocator described above. In this section we deal with the following problem: Given: The values of $p$, $l_1$, $l_2$ and the normalized loads $\lambda_1$ and $\lambda_2$, find: The delay probability distribution function and the cell loss probability for both service classes.

2.1 Approximate analysis

Let $P_{eff}(i)$ define the average normalized time-resource capacity allocated to service class $i$, $i = 1, 2$. Otherwise defined, $P_{eff}(i)$ is the effective probability with which a time slot is allocated to class $i$ at each instant. These probabilities can be approximately given by the equations:

$$
P_{eff}(1) = P_1(0) + (1 - P_2(0))p
$$
$$
P_{eff}(2) = P_2(0) + (1 - P_1(0))(1 - p)
$$

where $p(1 - p)$ is the time priority of service 1 (2) (i.e. the probability of getting a slot when contending with service 2 (1)) and $p_i(0), i = 1, 2$, the probability that no cells of class $i$ are waiting in the queue.

The above definition helps in splitting the system into two loosely-coupled queues with geometric arrivals and a geometric service time (Geo/Geo/1). The stationary state probabilities for the chains are given by:

$$
P_i(j) = \begin{cases} 1 - \frac{\lambda_i}{P_{eff}(i)} & j = 0 \\ \frac{P_i(0)}{1 - P_{eff}(i)} \rho_i^j & j > 0 \end{cases}
$$

where

$$
\rho_i = \frac{\lambda_i(1 - P_{eff}(i))}{(1 - \lambda_i)P_{eff}(i)}
$$

By combining 1 and 2 after some mathematical manipulations we get:

$$
P_1(0) = \frac{\beta^2 + 4p(1-p)(1-(\lambda_1+\lambda_2))}{2p}
$$
$$
P_2(0) = \frac{-\beta^2 + 4p(1-p)(1-(\lambda_1+\lambda_2))}{2(1-p)}
$$

with

$$
\Delta = \beta^2 + 4p(1-p)(1-(\lambda_1+\lambda_2)), \quad \beta = p(1-\lambda_1)-(1-p)(1-\lambda_2)
$$

The Complementary Probability Distribution Functions (CPDF) of the delay and the queue length can then be calculated:

$$
\log CPDF_d(d_i) = d_i \log \frac{1 - P_{eff}(i)}{1 - \lambda_i}
$$
$$
\log CPDF_Q(q_i) = \frac{\lambda_i}{P_{eff}(i)} + q_i \log \rho_i
$$

with $\rho_i$ given by 3 and $P_{eff}(i)$ by 1.

Evaluation of the Cell-Loss Probabilities

Assuming that the total capacity $L$ is divided into parts $l_1$, $l_2$ for the two service classes, the cell-loss probability for each individual service class $i$ is expressed approximately as

$$
P_{loss(i)} = Pr\{q_i > l_i\} Pr\{q_1 + q_2 > L\} \lambda_1 \lambda_2
$$

or

$$
\log P_{loss(i)} = \log CPDF_Q(l_i) + \log CPDF_Q(l_1 + l_2) + \log(\lambda_1 \lambda_2)
$$

The $CPDF_Q$, corresponding to the total queue length occupied by both classes, can be exactly derived as explained in section 3.2.

2.2 Exact 2-D Modelling and Numerical Solution

Let $n_1(m)$, $n_2(m)$ denote the number of cells of services 1 and 2 respectively, that are queued (waiting) in memory at time slot $m$. The vector $(n_1(m), n_2(m))$ is a homogeneous Markov chain with a state transition diagram as depicted in figure 2. Allowable states of this chain are those for which $n_1 + n_2 \leq L$, where $L$ is the assumed buffer length. By arranging them in a one-dimensional state vector and forming the corresponding transition matrix $P$ we can derive the equilibrium equation

$$
\pi = \pi P
$$

The trivial way of solving 9 is through using the recursive formula

$$
\pi(n) = \pi(n - 1)P
$$

until the Euclidean distance of any two vectors $\pi(n), \pi(n + 1)$ is less than a suitable stopping value. Alternatively, more sophisticated methods for solving 9 can be used, based on the special form of the matrix $P$.

Evaluation of the Cell Loss Probabilities

Apparently, cell losses can occur only when the system is found in states $(i, j)|i+j=L$, and cells of both services arrive. The cell loss probabilities for both services are calculated using the formulas:

$$
P_{loss(i)} = [\pi(l_1, l_2)(1-p) + \sum_{n=l_1}^{L} \pi(n, L-n)] \lambda_1 \lambda_2
$$
Figure 2: State transition diagram of the 2-D Markov chain

\[ P_{\text{loss}}(2) = \left[ \pi(t_1, t_2) \right] p + \sum_{n=t_2+1}^{L} \pi(L-n,n) \lambda_1 \lambda_2 \]  

Evaluation of the Delay PDF

The delay of a cell is measured with the number of slots between the arrival and the servicing of the cell. In order to find the delay PDF for service 1 for example, we assume that source 1 produces one test cell and then it is deactivated, while source 2 continues producing cells with the constant rate \( \lambda_2 \). By allowing \( k \) transitions beginning from the equilibrium state, the delay PDF for service 1, defined by:

\[ \text{PDF}_{D_1}(k) = \Pr\{\text{Delay}1 \leq k\}, \]

is given by the probability of finding the system at the state \((0,n_2)\), \( n_2 = 0,1, \ldots, L \), i.e.

\[ \text{PDF}_{D_1}(k) = \sum_{i=0}^{L} \pi_k(0,i) \]  

( \( \pi_k \equiv \pi \) after \( k \) transitions from the equilibrium).

A simple change of indices leads to the calculation of service 2 delay PDF. The closed-form solution as well as simulation results have confirmed the validity of our reasoning.

3 Resource Dimensioning and Allocation

In this section the problem of resource dimensioning and allocation is solved. It can be stated as follows:

Given: the absolute, average loads \( a_1 \) and \( a_2 \) (in cells/second) and the desired QoS parameter values, i.e. a certain delay percentile \( (d_i, CPDF_{D_i}(d_i)) \) and the cell loss probability \( (P_{\text{loss}}(i)) \), for each service class.

Find: the minimal capacities \( C \) and \( L \) for the time and the space resource respectively, as well as the allocation parameters (priority levels), \( p_i \) and \( I_i \), assigned to each service class

in order to guarantee the required service quality figures.

The solution of the problem requires the simultaneous determination of four different values \((C, p, L, I_i)\). However, a decomposition of the problem, which considerably reduces algorithmic complexity, is possible due to the following:

- The required buffer length, \( L \) is uniquely determined, given the probabilities \( \lambda_i \), \( i = 1,2 \) and the specified total packet loss probability.
- As packet loss probabilities are anyway small, buffer allocation has practically no impact on the delay figures and hence, it can be calculated independently of \( C \) and \( p \).

In the following, we first calculate the optimal values of \( C \) and \( p \) based on the delay requirements. Then, we present the closed form equation giving the required buffer length \( L \). Finally \( I_i \) is determined to guarantee the required loss probability for each service class.

3.1 Calculation of \( C \) and \( p \)

The approximate solution

Let \( t_i \) denote the slope of the line describing the behaviour of the delay CPDF, as derived by equation 5, i.e.

\[ t_i \equiv \frac{\log CPDF_{D_i}(d_i)}{d_i} \]  

and \( T_i \equiv e^{t_i} \) (14)

Then, eq 5 gives

\[ CP_{\text{eff}}(i) = C(1-T_i) + T_i a_i \]  

(15)

By combining equations 1, 2 and 15, after some mathematical manipulations we get \( C \) as the positive root of the second-order equation

\[ T_2 (1-T_1) - a_1 T_2 - T_1 C^2 + (T_1 T_2 (a_1 + a_2) - T_2 a_2 - a_1 T_1 a_1 T_2 a_2 + a_2 ) C - T_1 T_2 a_1 a_2 = 0 \]  

(16)

and \( p \) given by

\[ p = \frac{T_1 a_1 - T_2 a_2 + a_1}{a_1 + a_2} + \frac{T_2 - T_1 C}{a_1 + a_2} \]  

(17)
Returning to equations 5 and 14, we notice that the delay parameter $d_i$ is expressed in time slots. However, the delay is usually specified in absolute time units, so that the $d_i$ is not known a priori, because it depends also on the value of $C$. We can use, however, an approximate value for $C$ ($C > a_1 + a_2$) to get $d_i$ in slots and, if the solution is apart from the initial assumption, we can reassess it with a new value for the $d_i$. In any case, the relationship between $d_1$ and $d_2$ is exact, independently of the units assumed.

**Numerical approach**

Alternatively, instead of specifying $C$ (which depends on the absolute values of loads), we specify the load factor:

$$\rho = \frac{a_1 + a_2}{C} \quad 0 \leq \rho \leq 1$$

Let $D_1$ and $D_2$ be the delay CPDF for services 1 and 2 respectively. Both $D_1$ and $D_2$ are functions of 2 variables, $p$ and $\rho$. This pair of functions establishes a mapping, denoted by $w$, of the $(p, \rho)$ plane to the $(\log D_1, \log D_2)$ plane. On the later plane delay specifications are represented with a point which we will call "specification point". We wish to determine a point on the $(p, \rho)$ plane, that lies within the square of figure 3, is mapped to the specification point and has the maximal value of $\rho$.

According to the following observations, there is a unique such point:

- for constant $\rho$, $D_1$ is a decreasing function of $p$
- for constant $\rho$, $D_2$ is an increasing function of $p$
- for constant $p$, $D_1$ and $D_2$ are both increasing functions of $\rho$

The locus of points with constant delay on the $(p, \rho)$ plane will have, therefore, the form shown in figure 3. The region of acceptable QOS is represented by the shaded area, while the intersection of the two curves is mapped to the specifications point.

An iterative algorithm has been developed which starts from an arbitrary point $(p_0, \rho_0)$ and proceeds successively towards the optimal point by minimizing a suitable metric function on the $(p, \rho)$ plane. At each point $(p, \rho)$ the delay CPDF for each service class, required for the calculation of the distance from the optimal point, is derived by using the analytical tool presented in section 2.

### 3.2 Evaluation of the Optimal Buffer Length $L$

As can be derived from the state transition diagram of figure 2 the process that describes the total number of packets in the system, i.e. $(n_1 + n_2)$, is a 1-dimensional Markov chain whose equilibrium state probabilities can be easily evaluated using the conservation-of-flux laws. The total cell-loss probability can be expressed by the following equation

$$P_{loss} = P_{loss(1)} + P_{loss(2)} = Pr(q_1 + q_2 > L)\lambda_1\lambda_2$$

which can give the required total capacity $L$:

$$L = \frac{\log(P_{loss(1)} + P_{loss(2)}) - \log(\lambda_1\lambda_2)}{\log(1 - \lambda_1)(1 - \lambda_2)}$$

### 3.3 Buffer Allocation

**Approximate solution**

By using 7 the ratio of the individual loss probabilities is expressed as

$$\frac{P_{loss(1)}}{P_{loss(2)}} = \frac{CPDFq_1(\lambda_1)}{CPDFq_2(\lambda_2)}$$

or, by using 6 and substituting $l_2$ by its equal $L - l_1$,

$$l_1 = \frac{1}{m_1 + m_2} (m_2 L + \log p_{loss(1)} + k_2 - k_1)$$

with the constants $k_i, m_i, \quad i = 1, 2$ defined by:

$$k_i = \log \frac{\lambda_i}{P_{eff(i)}} \quad m_i = \log \frac{\lambda_i(1 - P_{eff(i)})}{(1 - \lambda_i)P_{eff(i)}} \quad , i = 1, 2$$

If, for a given operation point, $(\lambda_1, \lambda_2)$, the existing buffer length, $L$, is more than enough to guarantee the required $P_{loss(1)}$ and $P_{loss(2)}$, the buffer allocation problem can also be solved by using 21.

**Numerical approach**

In the same way as with the determination of $C$ and $p$ explained in the previous section, an iterative algorithm can be used to evaluate the optimal sharing of the available buffer space $L$.

### 4 Call Acceptance Control (CAC)

The CAC problem concerns the decision of whether a new call can be established under its required QOS, without violating the QOS limits of existing calls. In the case of our simple multiplexer we seek a region $R$ of the $\lambda_1 - \lambda_2$ plane, such that for every point $(\lambda_1, \lambda_2) \in R$, we can find a set of allocation parameters, $p, (1 - p), \lambda_1, \lambda_2, L - l_1$ which satisfy the desired QOS for both service classes. As far as the establishment of a new call does not bring the point outside $R$, it can be accepted. If we denote by $R_d$ the region of acceptable loads from the delay point of view and $R_l$ the one which can
satisfy the cell-loss requirements, then \( R = R_d \cap R_t \).

In a well-designed system, the space resource is dimensioned in a way that all the points inside the region \( R_d \) should also satisfy the cell-loss criteria. Otherwise stated, the buffer length should be large enough to guarantee the cell-loss requirements, for all combinations of loads which are acceptable from the delay point of view, i.e. \( R_l \geq R_d \) and, consequently, \( R = R_d \). This is so, because the time resource is usually the most expensive one. Under this assumption the only requirement for CAC is to keep the load point inside \( R_d \).

The way in which this requirement is met is clarified in the resource dimensioning and allocation paradigm which is presented in the following section.

5 Resource Dimensioning and Allocation Paradigm

Suppose that we want our multiplexer to operate all over a region of load points, as shown in figure 4 (a), with the following QOS requirements:

- **class 1** \( CPDFD_1(9\mu s) = 10^{-3}, P_{loss(1)} = 10^{-6} \)
- **class 2** \( CPDFD_2(4\mu s) = 10^{-3}, P_{loss(2)} = 10^{-9} \)

We seek the minimal dimensions of the resources, \( C, L \), which can guarantee the required QOS, as well as the corresponding CAC region on the load plane.

Approximate solution

Let assume a value of 1 Mslot/sec for the line capacity \( C \), which implies the values of 9 slots and 4 slots, respectively, for the delays \( d_1 \) and \( d_2 \) respectively. Using the analysis developed in section 3, we determine the locus of points on the load plane which exactly meet the above delay requirements. The locus \( C_l \) is shown in figure 4 (b). In the same figure the specified operation region (enclosed by the trapezium \( T_1 \)) is shown. One can observe that the assumed capacity \( C \) is not optimal. By letting the trapezium to adjoin the \( C \) curve, and using simple similarity rules we can compute a new value for \( C \), equal to about 930 Mslots/sec. With this new value, the delay figures corresponding to the specified percentiles are reassessed as \( d_1 = 8.36 \) slots and \( d_2 = 3.72 \) slots. The new curve \( C \) almost adjoins the operation region, as shown in figure 4 (c), hence the assessed value \( C = 930Kslots/sec \) is accepted. For each load point \( (\lambda_1, \lambda_2) \in C \) we calculate the required buffer lengths \( l_1 \) and \( l_2 \) which guarantee the loss probability requirements. The maximum total buffer length required is found to be \( L = 13 \).

In the case the multiplexer operates at a point with less requirements than the maximum length \( L \), the latter is also allocated according to eq 21. For example, if \( \lambda_1 = 0.2 \) \( \lambda_2 = 0.4 \), with \( p = 0.25 \), then the optimal allocation of \( L \) to guarantee \( P_{loss(1)} = 10^{-3} \) is found to be \( l_1 = 6.10 \) (6) \( l_2 = 6.90 \) (7). We can see from figure 5 that also the simulation results are in good agreement with what we initially specified and analytically found.

6 Conclusion

The problem of resource dimensioning, resource allocation and call acceptance control in a cell-level statistical multiplexer has been investigated. Two service classes, with different QOS (Quality-Of-Service) requirements (delay jitter and cell-loss probability) are multiplexed according to a sharing scheme with priorities. Specifically, under contention conditions, the outlink bandwidth (time resource) and the buffer space (space resource) are allocated to the two service classes according to certain allocation parameters (or priority levels), \( p_i \) and \( \frac{1}{l_i} \), \( i = 1, 2 \), respectively.

By introducing the notion of the effective capacity used by each service class, the two-dimensional problem is approximated by a model consisting of two loosely-coupled Markov chains with a closed-form solution.

The analysis leads to the calculation of the delay probability distribution and the cell-loss probability, when given are the dimensions of the multiplexer (line capacity, \( C \) and buffer capacity, \( L \)) along with the corresponding resource allocation parameters, \( p_i \) and \( \frac{1}{l_i} \).

The dimensioning problem is also solved analytically. It consists of determining the minimal values of \( C \) and \( L \) which guarantee the required QOS, expressed as a certain delay percentile and a certain probability of cell loss for each individual service class. The call-acceptance-control problem was also encountered through specifying an appropriate region on the load plane for which the specified QOS requirements for each class are satisfied: As far as any new call does not shift the operation point outside that region, it can be accepted.

The attractiveness of the approximate solution, lies in the fact that results can be derived by using even a pocket calculator. The derived formulas are simple and self-contained and, therefore, very useful for on-line call acceptance control and cell-level resource allocation.

In an alternative approach, the same problems have been answered accurately through a two-dimensional modelling. A numerical solution was obtained by developing a dedicated numerical tool (matrix manipulations, iterative dimensioning etc), suited mainly for dimensioning and off-line analysis.

REFERENCES

Figure 4: Multiplexer load plane: (a) Operation region (b) $R_d$ with $C = 1M$ slots per sec (c) $R_d$ with $C = 930K$ slots per sec

Figure 5: Delay and Queue-length Complementary Probability Distribution Functions, resource dimensioning and allocation paradigm

Strategies in Wide-Band Integrated Networks," presented at INFOCOM '84, San Francisco, CA, April 84.


