PERFORMANCE EVALUATION OF A DISTRIBUTED RESERVATION TDMA TO ACCESS AN ATM USER-ORIENTED SATELLITE SYSTEM

Stefano Agnelli, Vittorio Trecordi

Centro CEFRIEL
via Emanuelli, 15 - 20126 Milano - ITALY
Tel +39-2-66.100.750 - Fax 66.100.448

Abstract
This paper examines a distributed reservation TDMA algorithm for dynamic handing of bursty traffics in an ATM multiservice User-Oriented satellite network. Major aspects to achieve efficiency, robustness and fairness are highlighted. A performance evaluation carried out by means of an accurate simulation model is presented: a key issue is the adoption of a two-state Markov representation for bursty sources. Simulation results show the advantages (in terms of bandwidth usage) of the above access technique with respect to a possible static peak bit-rate bandwidth assignment.

1. Introduction
The growing demand for broadband services and the rapid advances in key technologies have fostered the definition of Broadband Integrated Services Digital Networks (B-ISDN). Asynchronous Transfer Mode (ATM) is currently recognized as the target transfer mode for supporting all services in a B-ISDN environment, providing switching and multiplexing facilities. Recent studies [Ku1] have focused on the introduction of ATM techniques in the world of satellite systems. Advanced features of regenerative multispot satellite with On-Board Switching [Co1] have made the application of this transfer mode viable in space communications.

A User-Oriented (UO) satellite system has been conceived to support directly attached single users with different arrangements, such as small terminals or LANs, belonging to a multiservice business communication environment. In this scenario the satellite constitutes the user network interface and has to support the variability and the burstiness of the traffics. An ATM multispot regenerative satellite can easily accommodate the different traffic profiles of various services and provides the UO system with full connectivity among all stations in an efficient and flexible way.

In the earth segment of the UO ATM multispot scenario, a set of up-link and down-link frequencies is assigned to each spot and each carrier is organized in a framed/slotted channel shared by a number of stations [Ex1]. The amount of bandwidth per channel is small (a few Mbit/s) with respect to the ATM terrestrial links because of the reduced power and antenna diameter required to keep low the user equipment cost. This context requires an appropriate choice of the multiple access protocol, which strongly impacts the network performance, expressed in terms of access delay versus traffic load [Jo1]. Random access techniques such as ALOHA are undesirable because of their large sensitivity to traffic load. In the frame of reservation TDMA techniques the application of a distributed access protocol seems very promising [De1][De2] since it achieves low access delay with a reasonable implementation complexity.

This paper focuses on a bandwidth assignment algorithm for handling bursty traffics with a distributed reservation TDMA technique. It also evaluates the performance of the proposed technique in terms of cell access delay by means of a simulation model. This model improves the single cell Poisson arrival model used in [De2], introducing the more realistic characterization of bursty sources based on a two state Markov chain, and offers a rigorous analysis of each contribution forming the overall delay.

We present in Section 2 some remarks on ATM services and access techniques. Section 3 introduces the key aspects on the access technique. Section 4 explains the algorithm application. Finally in Section 5 the simulation model is presented and engaged results are discussed.

2. Static and dynamic call handling
ATM is a connection-oriented technique: it requires that a virtual circuit is set up before starting user information transfer phase. ATM must offer an information transfer accommodation to both continuous and bursty services with very different traffic profiles and performance requirements. We investigate a reservation TDMA scheme to regulate the access of these services to the up-link frame in order to avoid collisions [Ja1][Ku1]. Reservations for continuous services are performed once per call at the virtual circuit set up, and where successful they result in a static slot assignment to each call (static call handling). As regards bursty services, slot reservations can be conveniently accomplished more than once during the call, on a per burst basis (dynamic call handling). The latter solution achieves a very efficient usage of the available bandwidth exploiting the statistical multiplexing of non-continuous sources and coping with their fluctuating resources demand. The unpredictable random access delay, to be added to the inevitable delivery lag due to the reservation, prevents from handling bursty traffics with real-time requirements in such a dynamic fashion. In fact these traffics necessitate a per-call reservation technique, which permanently assigns to each call a dedicated capacity whose bit-rate fits service peak bit-rate. Therefore, according to their tolerable delay, services in an ATM network are classified in [De2] as either non-isochronous or isochronous. The first class is made of bursty services which are delay-tolerant, such as file transfer, electronic mail, facsimile, data browsing and document retrieval. The second class is made of both bursty services which have real-time requirements and
continuous services, such as telephone, video-telephone, videoconference.

Satellite access can be achieved either through a centralized or a distributed technique. In the first solution each station addresses its reservation via satellite to a Network Control Center (NCC) which schedules the sharing of the channel among requests and notifies to each station its assigned slots. The whole operation roughly takes a double-hop propagation delay.

In a distributed access technique, reservations sent from each station are broadcasted by the satellite to the whole cluster of stations sharing an up-link channel capacity in the same spot. The transmission scheduling is determined through a local application of a cooperative algorithm, which prevents collisions in normal operation mode. This single hop solution approximately halves the delay of the centralized method.

As far as dynamic call handling is concerned, the reservation delay has to be engaged by each single burst. Therefore we focused on the distributed approach for uplink access of non-isochronous cells and addressed a bandwidth sharing algorithm based on concepts outlined in [De2]. Conversely we assumed that call set-ups of both isochronous and non-isochronous services are handled by the NCC, since the double-hop delay is negligible in the call set-up phase.

3. Algorithm key aspects

In an ATM network each source arranges the bit stream into the information field of fixed-size packets, called ATM cells. Each cell has a size of 40×384 bits (548 bytes), 40 bits for the header, 384 bits for the information transfer. The structure and the characteristics of up-link channel frame in the studied case are shown in Figure 1. Each control slot is permanently assigned to a station which uses this dedicated channel (Media Access Control) to transmit signalling and reservations. Each information slot carries 4 ATM cells.

The multiframe is defined as the set of successive frames which jointly contain the control slots assigned to all the stations sharing an up-link channel. This multiframe represents the basic structure of the distributed access algorithm of non-isochronous services. In fact the resource sharing lies in a sequence of contentions among reservations sent in a multiframe in order to acquire information slots available in a given successive multiframe. In the present study we consider a channel shared by 8 stations, each having its own associated control slot. Therefore each multiframe is made of two frames.

Each information slot, that is the minimum unit which reservations are issued for, carries a set of transmission units (i.e. ATM cells). A slot might be not entirely filled by the cells which it has been reserved for. This quantization is expected to have effects according to whether the unused capacity is exploited by other cells or not. In the first case quantization may result in a useful reduction of access delay, since some cells are transmitted before their reservations become effective (this phenomenon has been referred in [De2] as gate-crashing). In the second case an adverse wasting occurs. Both consequences grow when the number of cells per slot increases, while the overhead due to synchronization field becomes lighter.

The proposed number of ATM cells per slot has been chosen [De2] by trading off all the above factors.

As far as robustness is concerned, the distributed application of the same algorithm calls for countermeasures able to restrict common losses of information due to collisions in case of local failure. A very simple method to achieve such a goal is keeping no memory of received reservations, either honored or not, once the algorithm contention has been performed. Therefore the beginning of each multiframe identifies a regenerative point at which the assignment process restarts itself. Thus stations are required to renew unsuccessful reservations.

The round-trip propagation delay, which roughly takes 250-280 second, according to the geographical position of each station, makes the reservations effective 9 multframes after their transmission. Namely slots assignment in a multiframe is carried out analyzing the reservations sent 9 multframes before. During this interval a station can send reservations caused by further burst generations. Since a FIFO sequence has been adopted for cell transmission in order to limit the maximum delay, a pending cell whose reservation has been discarded can exploit a slot captured through a successful reservation for a following cell. This misalignment may be spread to a number of subsequent cells. The release of the liaison between each cell and its own reservation, caused by the gate-crashing effect as well, results in a suitable rearrangement of access delay.

Finally let us consider fairness issues. Broadcasted reservations are exploited to accomplish the resource assignment. Since the relationship between each station and its control slot is static, reservations are received in an unchanging sequence. If demands were honored on a
simple first come first served basis, this sequence would introduce an implicit priority. In order to achieve a fair access, the algorithm applies a round-robin procedure among the stations before serving reservations.

4. Algorithm definition

The procedure of slots allocation among a group of geographically distributed stations transmitting in an up-link channel develops as a series of independent contention sessions, each taking a multiframe. In a session, each earth station establishes how many information slots are assigned to itself and their position in the multiframe, and the reservation amount to be channeled into the control slot.

Once resources of the multiframe \( M \) have been allocated to both isochronous calls and reservations from other stations which currently have a higher priority according to the round-robin scheme, a station can use the remaining slots, if any, and picks them enough to honor its own reservation transmitted in the multiframe \( M-9 \) (the lag for reservation analysis is 9 multiframes long). Conversely a station undergoes a total or partial rejection when resources cannot satisfy its demands. In the captured slots a station transmits cells from its buffer with a FIFO priority.

The reservation amount that a station ships in the multiframe \( M \) is the sum of:
- the number of new cells produced since last reservation opportunity
- the cells which a reservation was sent for in the multiframe \( M-9 \) and are transmitted within the multiframe \( M \), because of a rejection.

Due to misalignment, indeed, some cells can have already been transmitted when a refusal of their reservation is decided. Obviously these reservations must not be renewed. That is why a plain comparison between forwarded and honored reservations is not sufficient for formulating the further reservation amount.

5. Modelling and performance evaluation

An accurate model of the above described access technique, based on a queueing network, has been designed exploiting the advanced graphic features of the software package Q+ by AT&T (Fig.2). Both ATM cells and reservations are represented as packets travelling into the network.

One of the most important features of the model is the adoption of an appropriate bursty cell generator. In [De2] a single cell source with Poisson inter-arrival was adopted. This choice is not well suited for the most of ATM bursty sources. We adopted a source which interleaves active and silent periods, both independent and exponentially distributed. In the active period it generates bits at its specific peak rate \( P \). The probability density functions for active and silent time duration are respectively:

\[
 f_a(t) = \zeta_1 e^{-\zeta_1 t}; \quad f_s(t) = \zeta_2 e^{-\zeta_2 t} \quad (t > 0)
\]

Therefore we can unambiguously identify a bursty source through the three parameters:

- \( P \) = peak bit-rate [bit/s],
- \( A \) = average bit-rate [bit/s],
- \( B \) = mean burst length [bit],

where:

\[
\zeta_1 = \frac{P}{B}; \quad \zeta_2 = \frac{AP}{B(P - A)}
\]

Each source arranges the bit stream into the information field of ATM cells. Therefore each bursty source is alternatively identified by the parameters:

- \( p \) = peak cell generation rate [cell/s],
- \( a \) = average cell generation rate [cell/s],
- \( b \) = mean burst length [cell].

The values \( p,a,b \) can be obtained dividing the values \( P,A,B \) by the number of bits in the information field of a cell. In the following we use indifferently \( P \) or \( p \), \( A \) or \( a \), \( B \) or \( b \).

In order to assess the performance of the distributed access a set of simulations has been carried out with the above model. The net channel capacity is 7168 cell/s, that corresponds to a net information transfer capacity of \( C=27752,512 \) bit/s, excluding cell header. \( N \) active earth stations (up to 8) are supposed to offer to the network a traffic load equally distributed among them. A portion \( C_i \) [bit/s] of this capacity is assigned to the isochronous calls, fitting their required peak bit-rate. The remaining channel capacity \( C_{ni} \) [bit/s] \((C_{ni}+C_i=C)\), which we refer hereafter as dynamic channel, is contend by \( N \) homogeneous non-isochronous calls (one per station) with a per-burst reservation. The number of calls in progress remain unchanged during each simulation, that is no tear-down and no further call set-up occurs. \( C_i \) and \( C_{ni} \) also remain unchanged during the observation period.

Each simulation focuses on the activity of non-isochronous calls, which are statistically multiplexed in the dynamic channel. The parameters identifying non-isochronous sources are the peak bit-rate \( P=192 \) and \( 768 \) kbit/s), the average bit-rate \( A=48 \) and \( 192 \) kbit/s) and the mean burst length \( b=10 \) and 100 ATM cells).

We define the dynamic channel load \( \rho_{ni} \) as the ratio between the total average non-isochronous traffic and the dynamic channel \( C_{ni} \):

\[
\rho_{ni} = \frac{NA}{C_{ni}}
\]

The parameter \( \rho_{ni} \) assumes the values 33.3%, 50.0% and 86.6%.
The list of the simulations obtained considering the combination of the above parameters is shown in Figure 3. Shadowed cases have not been developed since the peak rate $P$ of a single source exceeds the dynamic channel capacity $C_{nl}$. We collected statistics relevant to both the cell access delay $\delta$ and the capacity waste $\alpha$ due to quantization. Cell access delay $\delta$ is measured as the sojourn time in the queue representing the station transmission buffer, whose size is set infinite (no cell losses). Capacity waste $\alpha$ is the difference between the information transport capacity of the slots which have been reserved and their throughput, divided by the transport capacity of the reserved slots. Each run has been repeated up to 10 times with different seeds of the random number generator to achieve a satisfactory confidence level. The 95% confidence intervals were obtained assuming a T-Student distribution of results. The whole set of simulations takes about 360 hours of CPU time on a SUN 3/60 workstation. Figures 4, 5, 6, 7, 8, 9 show the average access delay and the maximum delay of the 90% and 99% of cells. For values higher than 99% the confidence interval is not thin enough, that is results have an inadequate reliability (error sometimes exceeds 10%).

A brief analysis of the sensitivity of the observed quantities led us to the following remarks:

- The burst length seems generally the main element affecting the cell delay. Average access delay becomes roughly double or triple when $b$ increases from 10 to 100 cells (see Fig.3).
- The capacity waste $\alpha$ is obviously lower where the burst length is much larger than the number of cells in a slot (b=100, Fig.3); in fact the quantization effect on reservation touches a smaller quota of cells in each burst.

<table>
<thead>
<tr>
<th>P</th>
<th>A</th>
<th>b</th>
<th>N</th>
<th>$\rho_{nl}$</th>
<th>$\delta$ ave</th>
<th>$\alpha$ ave</th>
<th>$C_{nl}$</th>
<th>$C_{nl}^{Pr}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>768</td>
<td>192</td>
<td>10</td>
<td>4</td>
<td>33.3</td>
<td>193</td>
<td>3.2</td>
<td>2304</td>
<td>3072</td>
</tr>
<tr>
<td>768</td>
<td>192</td>
<td>10</td>
<td>4</td>
<td>50.0</td>
<td>104</td>
<td>4.0</td>
<td>1536</td>
<td>3072</td>
</tr>
<tr>
<td>768</td>
<td>192</td>
<td>100</td>
<td>4</td>
<td>66.6</td>
<td>128</td>
<td>4.2</td>
<td>1152</td>
<td>3072</td>
</tr>
<tr>
<td>768</td>
<td>192</td>
<td>100</td>
<td>4</td>
<td>50.0</td>
<td>285</td>
<td>3.5</td>
<td>1536</td>
<td>3072</td>
</tr>
<tr>
<td>768</td>
<td>192</td>
<td>100</td>
<td>4</td>
<td>66.6</td>
<td>326</td>
<td>4.6</td>
<td>1152</td>
<td>3072</td>
</tr>
<tr>
<td>192</td>
<td>48</td>
<td>10</td>
<td>8</td>
<td>33.3</td>
<td>163</td>
<td>6.9</td>
<td>1152</td>
<td>1536</td>
</tr>
<tr>
<td>192</td>
<td>48</td>
<td>10</td>
<td>8</td>
<td>50.0</td>
<td>167</td>
<td>9.2</td>
<td>768</td>
<td>1536</td>
</tr>
<tr>
<td>192</td>
<td>48</td>
<td>100</td>
<td>8</td>
<td>66.6</td>
<td>192</td>
<td>9.4</td>
<td>576</td>
<td>1536</td>
</tr>
<tr>
<td>192</td>
<td>48</td>
<td>100</td>
<td>8</td>
<td>50.0</td>
<td>295</td>
<td>3.2</td>
<td>768</td>
<td>1536</td>
</tr>
<tr>
<td>192</td>
<td>48</td>
<td>100</td>
<td>8</td>
<td>66.6</td>
<td>339</td>
<td>7.6</td>
<td>576</td>
<td>1536</td>
</tr>
<tr>
<td>768</td>
<td>48</td>
<td>10</td>
<td>8</td>
<td>33.3</td>
<td>209</td>
<td>5.2</td>
<td>1152</td>
<td>6144</td>
</tr>
<tr>
<td>768</td>
<td>48</td>
<td>10</td>
<td>8</td>
<td>50.0</td>
<td>226</td>
<td>6.0</td>
<td>768</td>
<td>6144</td>
</tr>
<tr>
<td>768</td>
<td>48</td>
<td>100</td>
<td>8</td>
<td>66.6</td>
<td>66.6</td>
<td></td>
<td>576</td>
<td>6144</td>
</tr>
<tr>
<td>768</td>
<td>48</td>
<td>100</td>
<td>8</td>
<td>33.3</td>
<td>325</td>
<td>2.6</td>
<td>1152</td>
<td>6144</td>
</tr>
<tr>
<td>768</td>
<td>48</td>
<td>100</td>
<td>8</td>
<td>50.0</td>
<td>404</td>
<td>4.3</td>
<td>768</td>
<td>6144</td>
</tr>
</tbody>
</table>

\[\text{(*) error lower than 2.2% with confidence level 95%}
\]
\[\text{(**) error lower than 5.0% with confidence level 95%}
\]

---

Short-burst sources, which send reservations in a quasi continuous fashion, can achieve a low access delay with respect to the quarter-of-second propagation delay incurred by reservations. This desirable effect is due to the combined influence of misalignment and gate-crashing.

Average cell delay is almost insensitive to load in a wide range of values, which was already outlined in [De2] in a different traffic environment. This behavior is less appreciable where the ratio $P/A$ is high (Fig.6,9). According to the typical reservation TDMA behavior, it comes straight forward to think that the curve of access delay will remain flat considering dynamic channel load values below 33.3%. Because of this widespread insensitivity a fine network load control can be avoided and a simple rule for cell acceptance by NCC can be established.

Maximum delays are more sensitive to load than the average delay. This sensitivity is noticeably higher in the case of larger burst ($b=100$, Fig.7,8,9).

Finally in Fig.3 the bandwidth $C_{nl}$ used in the dynamic assignment is compared with the bandwidth $C_{nl}^{Pr}$ needed for the same traffic in case of a static peak rate assignment. The indubitable advantages of a statistical multiplexing of non-isochronous traffics have to be traded off with the originated delay.

**Summary**

A bandwidth assignment algorithm for handling ATM bursty traffics with a distributed reservation TDMA technique in a User-Oriented satellite system and a simulation model for evaluating its performance have been considered in this paper. The large instantaneous traffic variability in a UO scenario calls for a flexible management of the bandwidth. The effectiveness of a statistical multiplexing of traffics in an uplink frame with respect to a static channel sharing based on peak rate assignment has been outlined.

The studied algorithm accomplishes a suitable rearrangement of the access delay among cells; services have been shown to achieve somewhere an access delay lower than the quarter-of-second lag required for reservations propagation. Source average burst length has been evidenced to be the main element affecting the cell access delay. Delay performance has been verified to be almost insensitive to channel load variation in a wide range of values. Further investigations are suggested in order to ascertain whether a larger number of the information slots in a multiframe can smooth the impact of burst length.

**References**


---

Figure 4

Figure 7

Figure 5

Figure 8

Figure 6

Figure 9

5.1.5