ATM, HSLAN, and Images

B. Cousin, R. Castanet, L. Kamoun
LaBRI and IXL
351 Cours de la Libération
F-33405 Talence cedex

Abstract
This article presents the requirements of image transmission and the asynchronous techniques used in Broadband Integrated Services Digital Networks (B-ISDN) as well as in High Speed Local Area Networks (HSLAN). We examine the Asynchronous Transfer Mode (ATM) and the protocol of the Fiber Data Distributed Interface (FDDI) in detail. While the asynchronous mode of the FDDI protocol insures that the transmission of images conforms to time constraints, ATM proposes only a statistical answer.

1. Introduction

The growth of the formal exchange of information and the emergence of digital images as communication medium demand more and more efficient means of telecommunication. This efficiency must take into account the increased load, both in the number of communications and in the quantity of information exchanged. To take a familiar example, an telephone conversation generates only a few kilobit/s (from 4 to 64Kbit/s) whereas television transmission produces a thousand times more (from 2 to 200Mbit/s) [Guichard 90].

We can distinguish two major classes of telecommunication networks for the transmission of images: local area networks and wide area networks. In view of the heavy load generated by High Definition TV, only the most advanced techniques are feasible. For this reason, we are most interested in the asynchronous transfer mode (ATM) as representing the techniques of transmission over wide areas, and in the FDDI protocol representative of local area networks. Current technological breakthroughs including the use of fiber optic cable and the expected application range (about a hundred kilometers) make our comparison of these two techniques of transmission reasonable. This comparison should highlight the fundamental features of the two techniques of transmission, and accordingly it should permit us to demonstrate their adequacy for the transmission of images.

In the following sections, we present the criteria characteristic of the transmission of images. Then we study the new techniques of transmission, comparing those proposed for High Speed Local Area Networks (HSLAN) with those proposed for Broadband Integrated Services Digital Networks (B-ISDN). Finally, we study the adequacy of services provided, on the one hand, by the protocol FDDI and, on the other, by ATM, particularly with respect to the transmission of images.

2. Images

Digital images and digital voice have temporal constraints that we do not ordinarily encounter in conventional data transfer. These temporal constraints bind samples. A sample is that portion of a signal that is digitized. For example, a sample could be a group of bits, one byte of coded sound, or a line of an image. In order to express these temporal constraints, we need to distinguish two types of intervals of time: we denote Tij the interval of time between two samples i and j; we denote T0j the interval of time between the origin and a sample j (See figure 1). Clearly the first type of interval is relative, depending on the two samples, whereas the second type is absolute.

We indicate the moment of production of the sample by the sender (emitter) with the notation Te. Likewise, we use the notation Tr to indicate the moment that the sample arrives at the receiver. The production of a good movie requires that two constraints must be satisfied. First constraint: the delay after the sending of a film must be humanly tolerable, virtually instantaneous. We refer to the time that one must wait to see the first image of a movie as Tmax. This time is critical if the user intervenes in the unfolding of the movie; that is, if the film is in any sense interactive. Second constraint: the images should appear on the screen of the receiver at the same speed relative to one another as they are produced by the sender. If these two constraints are satisfied, then the film is received with temporal integrity. Two relations suffice to express these constraints:

(1) $\forall i\ Tr_{0i} < Tmax + Te_{0i}$.

(2) $\forall i,j\ Tr_{ij} = Te_{ij}$.

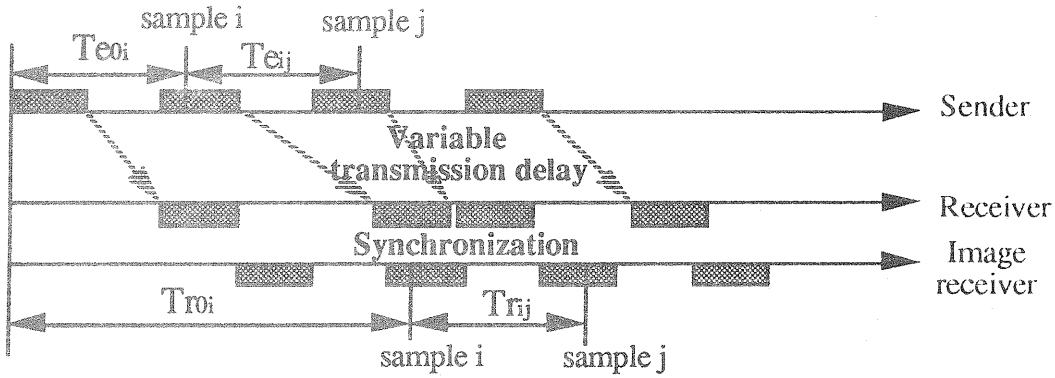

Figure 1 -Image temporal constraints

These temporal constraints exist only if the film should be visualized at its arrival on the receiver (in real time, no less!). These constraints do not exist if the film is broadcast in deferred time (for example, if it is pre-recorded for later broadcast), and if it is thus consequently stored on its arrival at the receiver. In such a case, the transmission of the film can simply be treated as the transmission of a large file.

In fact, in as much as they are located on distinct sites, the receiver of images is completely independent of the sender of images, and it is thus difficult to respect these two constraints.

Conventionally, the clock of the receiver of images is slaved to the clock of the sender by means of a synchronization included in the signal . Since the conventional methods of transmission use **synchronous** technique, the intervals of time between samples are preserved during their transmission. The synchronization of the receiver with the sender of the images is therefore easily achieved. It suffices to slave the receiver's clock to the flow of the received images. Only a constant delay is added.

The most current techniques of transmission now use **asynchronous** transmission. With this technique, the delay in the transmission of the samples varies: it depends on the method of media access, on the resolution of collisions, on the load, etc. Consequently the time separating two samples at their reception may differ from the time separating them at their transmission. We can no longer count on slaving the clock of the receiver directly on the flow of received images.

The technique of asynchronous transmission permits a better use of support than does synchronous transmission because sporadic flows can be compensated for. Asynchronous transmission works well with dynamic allocation methods of the bandwidth between different links as a function of the load. Nevertheless, in order to be efficient, the overhead introduced by this dynamic management must be compensated for by a better allocation of the traffic. In contrast, synchronous techniques use a method of static allocation that requires little or no management overhead.

If we use an asynchronous technique of transmission, we have to reconstruct any missing temporal information locally on the receiver by means of specific delimiters inserted explicitly in the signal at the time of transmission. But, if the transmission is asynchronous, the temporal reconstruction can be based only on the local clock of the receiver. Some drift and variations between the clocks of the sender and receiver will inevitably exist. By variations we mean fluctuations from the period of the clock around its nominal value. By definition, these variations has a nil average, whereas the effects of the drift between the two clocks are cumulative.

A buffer of units of transmission is adequate to obviate the variations in the local clock. If the variations of the clock conform to a symmetrical law, and if the variations are independent of one another, then a buffer of 40 units of transmission has an overflow probability that is lower than 10^{-6} and a buffer of 60 units has a probability lower than 10^{-9}.

Clock drift or variations between the receiver's clock and the sender's clock are inevitable. First of all, in fact, communications based on images generally last several hours, and this long duration makes even the slightest drift between the clocks quite noticeable. A deviation of 10^{-4} during a transmission of more than an hour introduces a delay of about 10 and more images. In view of the

transfer rate required for the broadcast of digital images, the quantity of information at stake is several megabits. Secondly, the greater the frequency of a clock, the greater is the accumulation of its drift from another clock, and we envisage a transfer rate requiring very high frequency.

Three methods exist for resolving the problem of clock drift. The first method, which we have already described, but which is not applicable in our application, consists of slaving the clock to the flow of data. The second method performs poorly under conditions of long duration and/or high rate of transmission. It entails the use of a buffer which is kept half-filled. This buffering permits us to adapt the flow of data to the clock of the receiver of images. This method has three drawbacks: it introduces great delay; it is costly in terms of memory; it is poorly adapted to transmissions of variable duration. In the third method, the drift between the clocks can be corrected if it is suppressed or if one inserts certain periods for the synchronization of the flow at reception. In either case, these added or suppressed periods can be handled either in the samples (or in the appropriate units of transmission) or in units designated for this purpose and insert deliberately in the transmission. The suppression or the insertion of units of information need not alter the quality of the image. In fact, they are usually physiologically tolerable if they are not too frequent. In any case, care must be taken to make either these insertions or these suppressions compatible with any techniques of compression or synchronization in order to facilitate good reproduction of the image.

This phenomenon of clock drift exists not only at the image receiver between the speed of reception and the speed of projection, but at each intermediate site (if they exist) with respect to the speed of reception and the speed of retransmission.

This method of correcting clock drift by insertion or by suppression necessitates the storage in memory of an entire image at the receiver. This buffering is used, first of all, in order to synchronize the local clock if it is more rapid than the received flow in displaying the preceding image a second time; secondly, in order to regenerate any part of the image missing or damaged in transmission; and finally, in order to absorb the variations in speed of transmission. Intermediate stations (if they are necessary), having no knowledge of the semantics of the transmit data, resolve the clock drift by inserting or suppressing some additional units of transmission, and they absorb variations by internal management of buffers of units of transmission.

In order to handle them in the optimal manner, both the loss of units of transmission and the creation of such units requires a means of identification of units of transmission. This identification facilitates the use of techniques for the recovery of missing or supplementary units of transmission working from the buffered image in the image receiver.

Two features are the deciding factors in the transmission of images: the protocol must guarantee that a maximum delay exists (i.e. that the delay of any given transmission is bounded), and it must guarantee a minimal throughput.

The first constraint is met if the network can guarantee that the maximum delay suffered in transmission will be smaller than Tmax for all samples of the signal. We note that tolerable waiting time is relatively great. A human being has a reaction time close to one-tenth of a second, when the networks that we envisage have an average delay of transmission close to a millisecond.

The second constraint is satisfied if, on the average, the throughput necessary for the transmission of the images of the movie is assured. The first constraint allows us to control the variance of this average. Memory buffers absorb the overload, that is to say, the moments when the immediate load is greater than the average load.

3. ATM

Asynchronous technique permits the development of an supple and all-purpose infrastructure of communication able to sustain very heavy loads. It involves a time multiplexed digital technique somewhere between packet switching used in data transmission networks and circuit switching used in telephone networks. As in the technique of packet switching, the frames hold a block of data information and an address which identifies the channel. The flow on each channel is arbitrary. As in the technique of circuit switching, we have no control over errors nor over fluctuations. The size of the frames is fixed, so the mechanisms for switching are simplified in the extreme. This enables very high speed switching devices to be built.

These two switching techniques give rise to the following four phenomena. First of all, our lack of control over errors produces a fault rate equal to the fault rate of the medium. The procedures and supports used nowadays sustain a particularly low rate of faults. Secondly, our lack of flow control may entail congestion at each switches. Global control of the network, nevertheless, allows the availability of required resources to be insured before the use of a new channel. The rate of loss due to congestion is thereby considerably reduced. Thirdly, in each switch, the waiting queues needed for robust switching of frames give rise to an increase in the delay of transmission. Small frames and the adaptation of the nominal flow on the network to the required service should reduce this growth of queues to manageable proportions. Fourthly and finally, the technique of asynchronous transmission inevitably introduces the phenomenon of variable delay so harmful for the transmission of periodic data.

Thus a well functioning ATM depends essentially on statistical considerations. Certain calculations prove that the rate of error and of loss, the influence of load on the rate of congestion and the variation in delay of transmission will remain within acceptable values for the applications that we envisage [Boyer 87].

The technique favored by local area networks takes the opposite approach from that of ATM. One chooses to share the same medium among all of the stations. This choice has as its first consequence the elimination of the intermediate devices--the switches--and the delays that they entail. Unfortunately, if the congestion at the switches disappears with them, then access to the medium--the resources shared by all the stations-- becomes critical. For this reason, local area networks involve specific methods of access to the media (MAC: Media Access Control).

4. FDDI

The protocol FDDI uses an access method called Token Ring. FDDI stations are connected in a ring where a token pass round. Any station that wants to transmit over the ring has to seize the token. At the end of its transmission, that station must relinquish the token. Every station, thereby, obtains the right to use the medium turn by turn.

The FDDI protocol enables transmission in synchronous mode. FDDI uses fundamentally a technique of asynchronous transmission (that is to say, the delay in transmission is variable), but this protocol stipulates two modes of transmission: the asynchronous mode and the synchronous mode. The synchronous mode guarantees a station a pre-allocated bandwidth and the right to transmit with an average periodicity equal to a value negotiated among all the stations. This periodicity is referred to as the Target Token Rotation Time (TTRT). Furthermore, the protocol guarantees a maximum rotation time of the token that cannot surpass 2xTTRT [Johnson 86].

At first glance, then, it seems easy to transmit voice or images by means of the protocol FDDI. Nevertheless, if we disregard the rate of transmission, which seems barely reasonable in conveying images in good condition, then we find that variations in load can lead to variations in the transmission delay of images.

During the initialization phase of the FDDI protocol, all the stations connected to the ring negotiate the value of TTRT. The TTRT chosen is the smallest. TTRT has to lie between the two values, Tmin and Tmax. The value Tmin corresponds to the minimum time for the management and rotation of the token. A TTRT value less than Tmin would not even allow the token to reach all of the stations, and thus such a time is unacceptable. A TTRT value greater than Tmax is conceivable without major problems except that it creates a partition of the medium somewhat prejudicial to equal access because a station holding the token could very well keep it a very long time. Furthermore, a TTRT value greater than Tmax slows the detection of errors and in addition the reconfiguration of the ring.

Let us note that the smaller the value of TTRT, then the more important becomes the amount of time dedicated to the management of access to the medium. In fact, the number of rotations of the token per unit of time is inversely proportional to TTRT. Thus the token consumes a great part of the bandwidth. We have every interest in sustaining as great as possible a TTRT within the limits of foreseeable applications.[Dykeman 88]

In order to insure a given flow to a channel sharing a connection with other channels, we can either adjust the frequency of access to the network or adjust the quantity of data sent during each session of access. Thus, in order to have a low frequency of access (i.e. a large TTRT), it is necessary to transmit a great deal of data simultaneously. We have, therefore, a tendency to group together several samples of the image.

The problems of data corruption, of synchronization, and of adaptation of the receiver to the flow of data--all oblige the receiving stations to buffer part of the images. Buffering at the level of the sender allows us to send great frames onto the network and thus necessitates only a long TTRT. In fact, the accumulation of these different bufferings is quite acceptable physiologically; a delay of several images occasions a delay of less than a tenth of a second, a negligible delay for a human being. We recall that this delay is applicable to all the images, and thus only the beginning of the film is affected by it. Nevertheless, this buffering does demand a great deal of memory .

Once the TTRT is fixed, once we know the average throughput di required by each station i to transmit the images in real time, we should limit the number of bits sent by each station at each session with the token in order to insure that we always maintain the following relation: (3) $\sum_i d_i \leq D$. That is to say, the sum of throughput sent should be lower than the effective throughput D of the network. This enables congestion of the medium to be avoided. The effective throughput is obtained by starting from the nominal throughput minus the throughput used to manage the network, essentially the packaging of the frames and the management of the token. If the protocol were perfectly synchronous, then station i would have to transmit li = di x TTRT bits every TTRT seconds. Network management has the responsibility for maintaining the statement (3). Every station requesting to transmit in synchronous mode calls the network management for a reservation of the average throughput required [FDDI 88].

Unfortunately, the load on the network can make the moment at which the token arrives at a station vary wildly (Recall that a station must capture the token before it can transmit). This moment is remembered by the token rotation timer (TRT) local to each station. It may be early or late with respect to the negotiated TTRT period. Logically, in order to maintain the inequality in relation (3), each station i should have the right to transmit at most dixTRT bits. This quantity is extremely difficult to manage because the TRT varies as a function of the load with each rotation of the token. Moreover, the implementation of FDDI does not permit us to get the value of TRT. We risk, then, exceeding the throughput di attributed to each station, and thus violating the inequality in relation (3), if we do not adapt the length of a frame to the rotation time of the token.

However, if the token is early, this indicates that the network is underloaded, and thus it is permissible to transmit di x TTRT bits. Inversely, if the token is late, then the mechanisms of the protocol FDDI insure that the delay cannot surpass 2 x TTRT, even if all of the stations transmit the entirety of their throughput synchronously. Moreover, the FDDI protocol requirement of updating the different token rotation timers (which control the period of rotation of the token) guarantees that the delay will not be cumulative. Accordingly, the prior and permanent recording of TRT in the receiver of the least li bits enables it to absorb the maximum delay.

In fact, the protocol is self-regulating because the overload induced by the token diminishes if the time between two passes of the token grows. Furthermore, if one of the stations does not use the entire throughput allocated to a synchronous transmission, then the unused time will be recovered, first of all, to insure other synchronous transmissions, in recovering the delay, and in re-establishing the negotiated frequency of rotation of the token; secondly and ultimately to authorize asynchronous transmissions. The control of the quantity of data transmitted at each capture of the token does not have to be managed at the level of the FDDI transmitter; the normal throughput of the sender of images naturally assumes this role.

In conclusion, we propose to use the synchronous mode of the FDDI protocol to transmit images of a movie. The average transmission rate di necessary to the transmission of the movie should be known, and the application should request to the network manager to make an appropriate reservation for the duration of the movie to guard against congestion of the media. The negotiation procedure for the TTRT could then be started, if required. The smaller the value of the required TTRT, the smaller will be the delay in transmission. However, we have already raised the idea that the efficiency of the FDDI protocol will be accordingly weakened. The calculations that we have undertaken indicate that the ideal value lies in the neighborhood of twenty milliseconds [Cousin 90].

Independently of the fact that the negotiated value of the TTRT should lie between Tmin and Tmax to insure the proper global functioning of the network, our application can accommodate a large range of values for TTRT. The application should be able to transmit at most li = di x TTRT bits at each rotation of the token.

Accordingly, at the price of a slight delay equal to TTRT due to the buffering at the level of the receiver of images, we realize that it is useless to request a rotation time equal to half of the delay of required transmission, as that would let it over-determine the maximum rotation time guaranteed to be more than 2 x TTRT.

5. Conclusion

In view of this study, we can observe that the two techniques of transmission, while on first glance presenting numerous dissimilarities (notably structural ones) resolve the same problems in a similar manner, namely in the preservation of the synchronization in the transmission or in the minimization of delays induced by intermediate equipment.

Yet the topology of a ring in the FDDI protocol both necessitates and allows controlled access to the medium, and favors the management of a method of access favorable to the transmission of images by the creation of two modes of transmission--both synchronous and asynchronous. The synchronous mode of transmission guarantees a station an average throughput and the right to transmit with a periodicity, on the average, equal to a value--the TTRT-- negotiated among all the stations. Moreover, this mode guarantees that the maximum rotation time cannot exceed 2 x TTRT.

The technique of transmission used by ATM makes it impossible to offer the same service. In effect, it were possible to pre-allocate a certain number of frames in a recurrent fashion on the entire route of links between the sender and receiver of images to convey the samples of the movie, but it is still impossible to avoid the occasional loss caused by congestion at the level of the ATM switches. Nonetheless, certain studies carried out for the transmission of voice have shown that the probability of congestion actually occurring (and thus the probability of loss) is extremely low.

We observe that in order to allow the transmission of synchronous information with asynchronous techniques, it is necessary to transmit temporal information explicitly; it is further necessary to supply sufficient buffer space in memory at the level of the receiver to accommodate the inevitable variations in delay in transmission. These memory buffers imply a systematic delay inhospitable to interactive applications. In short, the great throughput required by the transmission of images obliges one to use a great quantity of rapid access memory. Yet the memory buffers necessary at the level of the receiver in order to allow the use of an asynchronous technique of transmission can be usefully exploited to detect and then to correct loss, corruption, and duplication in all or part of the images of the movie. A complementary study is currently underway to evaluate the behavior of the FDDI protocol and the ATM technique for the transmission of images.

[Boyer 87] P.Boyer, J.Boyer, J.R.Louvion, L.Romoeuf, "Modelling the ATD transfer technique", Traffic engeneering for ISDN, Como, 1987.
[Cochennec 85] J.Y.Cochennec,P.Adam, T.Houdoin "Asynchronous time-division Networks : Terminal Synchronization for video and sound signals", GLOBECOM'85, 1985.
[Cousin 90] B.Cousin "Evaluation du protocole FFDI pour la transmission d'images", rapport de recherche LaBRI, à paraitre.
[Dykeman 88] D.Dykeman,W.Bux "Analysis and Tuning of the FDDI Media Access Control Protocol", IEEE Journal on Selected Areas in Communications, Vol 6 n°6, july 1988.
[FDDI 87] "FDDI Token Ring Media Access Control", ANSI X3.139, 1987.
[FDDI 88] "FDDI Station Management", ANSI X3T9.5, 1988.
[Filipiak 89] J.Filipiak "Structured systems analysis methodology for design of an ATM network architecture", IEEE Journal on Selected Areas in Communications n°7, 1989.
[Guichard 90] J.Guichard, G.Eude,"Visages", L'écho des Recherches n°140, 1990.
[Johnson 86] M.Johnson "Reliability Mechanisms of the FDDI High Bandwidth Token Ring Protocol", Computer Networks and ISDN Systems n°11, North Holland, 1986.
[Lam 78] S.S.Lam "A new measure for characterizing data traffic", IEEE transaction on communications, vol COM 24 n°1, january 1978.