Multicast IP support for distributed conferencing over ATM*

A. Azcorra[1], T. Miguel[1], M. Petit: [1]{aazcorra,tmiguel,mpetit}@dit.upm.es
L. Rodrígues[2], C. Acuña[2], P. Chas[2]: {lidia,carlosa,plchas}@tid.es
J. Bastos[3], V. Lagarto[3]: {bastos,lagarto}@cet.pt
1) Dpto. Ingeniería de Sistemas Telemáticos, UPM, Madrid, SPAIN
2) Telefónica Investigación y Desarrollo, Madrid, SPAIN
3) Centro de Etudos de Telecomunicações, Aveiro, PORTUGAL

June 17, 1996

1 Introduction

This article describes the construction of a TCP/IP network for the support of distributed multimedia conferencing on top of an international ATM network, and draws some general conclusions for the design of networks with similar requirements. An excellent tutorial describing the principles of internetworking over ATM WANs may be found in [**?**].

The network described in this paper was designed to support the second RACE Summer School that was held, during July 94, in a distributed manner among five sites: DIT-UPM (Spain), TID (Spain), CET (Portugal), U. Aveiro (Portugal) and Ascom (Switzerland). The five sites were interconnected by means of an ATM network, terrestrial and satellite, that provided user network interfaces at 155 Mb/s (STM1) and 34 Mb/s (E3). The devices used to access the ATM network included routers with ATM interface, ATM CSU/DSU equipment, and Ethernet IWUs for satellite ATM connections. Both AAL3/4 and AAL5 had to be combined because of compatibility with available equipment.

The ISABEL[**?**] application was used to support the RACE Summer School, requiring the transmission of digital video (M-JPEG compression), digital audio (8 bits, μ law with silence detection) and data. The different types of traffic generated by the application, and the various IWU devices used to access the ATM network imposed very particular requirements over the TCP/IP protocol architecture and configuration (e.g. it was necessary to handle in a homogeneous way both multicasting, for multimedia traffic, and unicasting, for reliable data traffic).

The next section of the paper describes the communication service requirements that were derived from the distributed multimedia application. Section 3 describes the telecommunication infrastructure and equipment that was available to build the network. The different approaches and alternatives to network design are described in Section 4, that has been divided in three subsections that focus on the design of the network at the IP layer, the mapping of IP over ATM, and the provision of multicasting service. Finally, Section 5 draws some conclusions intended to provide general guidelines for the deployment of other networks with requirements similar to the ones we had.

2 Service Requirements for Network Design

Five sites had to be connected by the network: DIT-UPM, TID, CET, U. Aveiro and Ascom. The

[1]This work has been partly supported by the CEU under projects ISABEL, BRAIN and IBER

traffic and service requirements were derived from the ISABEL application, which is a CSCW system that was installed at each of the sites to support the distributed conference.

The ISABEL application runs on a SUN station fitted with a Parallax video capture and compression board. The traffic generated by the different components of the application may be divided in three categories: conventional data, audio, and video. The multimedia capabilities of the application allow the transmission of one video+audio stream locally generated, and the simultaneous visualization of N video+audio streams received from remote applications. Each video stream is presented on the screen on its own window, while audio streams are digitally mixed to provide a single stream that feeds the loudspeaker.

Conventional data traffic required multipoint, reliable service. The need for multipoint traffic arises from the transmission of data and control information among the five sites. Application components such as white board, the distributed editor or the distributed pointer are examples in which the need for reliable, multipoint data traffic is clear. The aggregated required bandwidth was around 10 Kb/s.

Audio traffic required a point to multipoint, connectionless, unreliable service. The reason to select an unreliable service was the tradeoff between reliability and delay jitter. The reason to require a point to multipoint service is that sound produced at one site had to be received at all the other sites. The bandwidth required by each source was an intermittent constant bit rate stream of 64 Kb/s, generated by a 8 bits μ law conversor sampling at 8 KHz with silence detection and no compression. Audio traffic was extremely sensitive to delay jitter and data loss. Audio traffic sources were correlated: only one of them was active at any given moment (e.g. either the speaker, one person in the audience making a question, or one participant in a distributed panel).

Video traffic required a point to multipoint, connectionless, unreliable service. The reasons to define these service requirements were similar to the ones explained for audio traffic. The approximate bandwidth required by one source, at top quality, was around 2 Mb/s, generated by the Parallax board as an M-JPEG compressed stream at 13 frames per second and 800x600 pixels with true color. Contrary to voice traffic, video traffic was variable bit rate, and the 2 Mb/s represent the average, with larger peaks produced by more complex frames. Video traffic sources were correlated. The two most usual traffic situations were produced by either, one speaker in top quality and remote audiences in low quality (small images, low frame rate), or, two debating speakers in medium quality. These configurations required an aggregated average throughput around 3 Mb/s.

Absolute delay between any two endpoints should be kept low because we were dealing with interactive applications. Good quality is obtained with delay figures below the 200 millisecond level. Delay jitter should also be kept low to avoid discarding video frames or audio samples because of late arrival.

The service requirements previously described should be satisfied by constructing a network using the existing resources and equipment for the Summer School.

3 Available Infrastructure and Equipment

The ATM terrestrial infrastructure available for the network was composed by three ATM switches connected by two trunk links, one from TID to DIT-UPM and another from TID to CET. The University of Aveiro had an access link to the switch located at CET. The node at CET was lent by Alcatel for this experience, the node at TID is a product of Telefónica's research programme RECIBA, and the node at DIT-UPM was lent by project RAL-ATM from the Spanish research programme PLANBA.

As no terrestrial links were available to Ascom, EUTELSAT lent transponder capacity, and Marconi provided earth stations to connect CET and Ascom. The RACE-CATALYST project lent an interworking unit that provided ethernet emulation over ATM, which was transmitted using the satellite link. Unfortunately, it was not possible

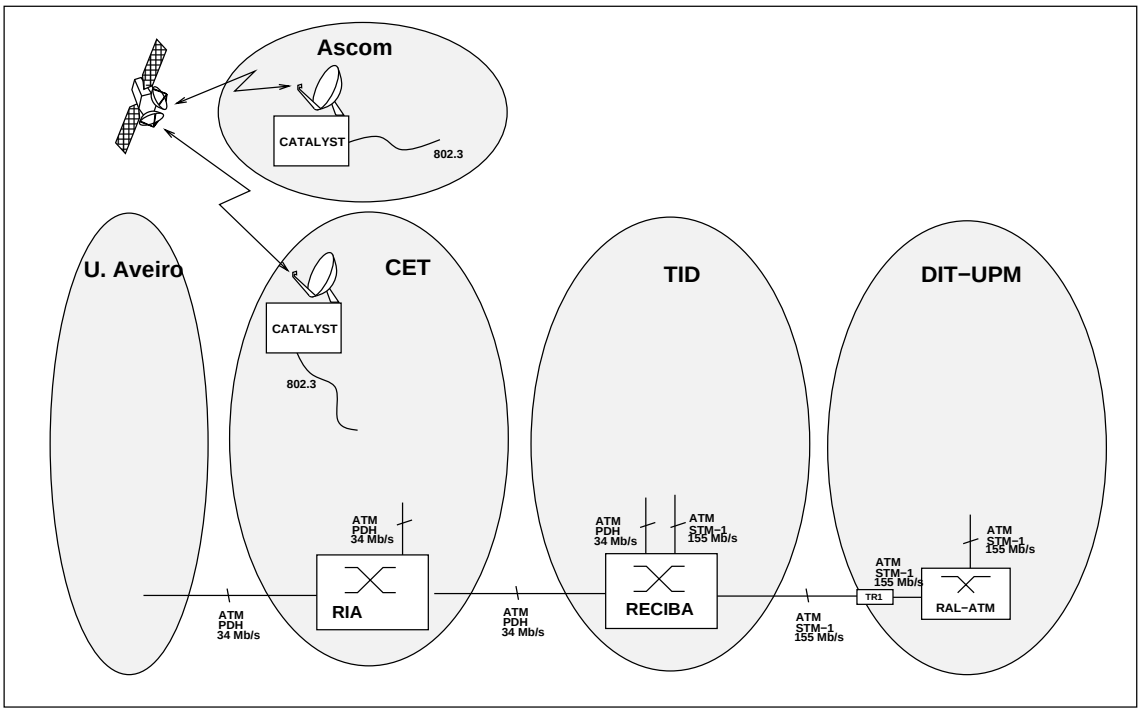

Figure 1: ATM and satellite networks

to provide native ATM access, and therefore we could not build an ATM network that spanned the five sites. As it is depicted in Figure 1, the available infrastructure may be summarized as an ATM network spanning CET, U. Aveiro, DIT-UPM and TID, plus an emulated ethernet network connecting CET and ASCOM.

Therefore, the connection between Ascom and the remaining sites was only possible by routing at the IP layer from the satellite ethernet to the IP network to be built over the terrestrial ATM network. Although the CATALYST system is very sophisticated, the ethernet service provided by it could only be treated as a physical ethernet, connecting a router placed at CET and the Sun workstation at Ascom. For this reason, this article will focus on the design of the IP service over the terrestrial ATM network.

The ATM access equipment available was not the same at every site. At TID, U. Aveiro and CET the equipment consisted in a Digital Link DSU, fitted with an E3 ATM interface to access the ATM network and an HSSI/DXI interface, plus a Cisco AGS+ router with HSSI/DXI, ethernet and FDDI interfaces. At DIT the access equipment was a Cisco 7000 with ATM, FDDI and ethernet interfaces.

The end systems to be connected to the IP network were SUN workstations. ATM cards and ethernet cards were available for the workstations. Some sites also had available FDDI cards.

The resources described in this section had to be arranged and configured in the appropriate way to provide the required service.

4 Design decisions and solutions adopted

The design process of the network is presented in three views. First, we present the design of the structure of the network in terms of independent IP subnets requiring IP routing to provide interconnection between them. Second, the different alternatives of building an IP virtual subnet over ATM are discussed. Finally, we explain how the multicasting service was provided both in the reliable and the unreliable modes.

4.1 Network Architecture at the IP Layer

The first design idea was to build a single IP virtual subnet over the ATM network, by direct interconnection of the end systems fitting each SUN workstation with an ATM card. This approach provides sufficient bandwidth and extremely low delay and delay jitter. Unfortunately, the available ATM cards only had drivers for Solaris, while the Parallax card only had drivers for SunOS. Moreover, all the ATM cards had SDH/STM-1 physical interface, while at some ATM switches the only available user-network interface was PDH/E3. Another problem that would have made not possible to use this approach was that our ATM cards could not perform ATM traffic shaping, and in spite of having all the capacity of the trunk lines dedicated for the Summer School, there was a bottleneck from the 155 Mb/s of the STM-1 trunk line to the 34 Mb/s of the E3 trunk line.

For these reasons this first design approach was discarded in favor of using the routers to access the ATM network. This implied that the end systems were linked to the routers by means of conventional LAN technology, either Ethernet or FDDI depending on each site, while the routers were interconnected through the ATM network. This second design would have resulted in a single IP virtual subnet by direct interconnection of the routers, plus DSUs, over the ATM network. This is a very nice design with good performance. The ATM 155-to-34 bottleneck is solved because the Cisco 7000 has some traffic shaping capabilities. The ATM physical interface problem is also solved because routers, plus DSUs, were compatible with ATM switches at each site. The problem we encountered was that the Cisco 7000 provided at that date IP encapsulation over AAL5, following RFC 1483[**?**], while the AGS+, plus Digital Link DSUs, provided SMDS over AAL3/4 encapsulation. The protocol architecture of both encapsulation schemes is shown in Figure 2.

To solve the incompatibility of IP encapsulation methods it was necessary to route at the IP level. To achieve this, a SUN station with an ATM card was installed as a router at TID. This SUN station was connected to the Cisco 7000 by means

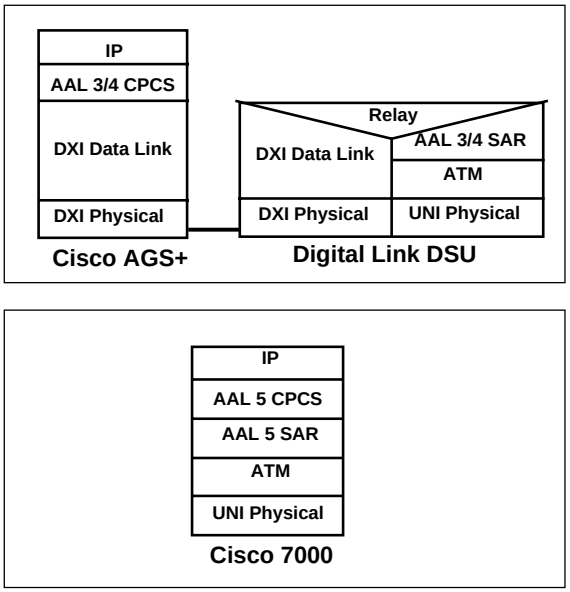

Figure 2: ATM encapsulation alternatives for IP

of the ATM link and to the Cisco AGS+ by means of a dedicated ethernet segment. This was the selected solution, depicted in Figure 3, that allowed end to end IP connectivity between all sites. The obvious implication is that we have now two IP virtual subnets instead of a single one, as it was initially intended. The functional connectivity is solved, but performance in terms of bandwidth, delay and delay jitter is degraded because of the Ethernet bottleneck and an additional routing hop.

4.2 Network Architecture of virtual subnets over ATM

After designing the network architecture in terms of IP subnets it remains to be discussed how each IP virtual subnet was to be built over ATM. The most advanced solution was considered to be building each virtual subnet over either an ATM LAN[**?**] or a connectionless service[**?**, **?**]. Both alternatives provide a flexible way to configure an IP virtual subnet over an ATM network. The ATM LAN alternative was not available at the time of the Summer School, but the connectionless service was feasible using the indirect approach.

In the previous subsection it was shown how it was neccessary to divide the network in two IP subnets, each of the using a different IP encap-

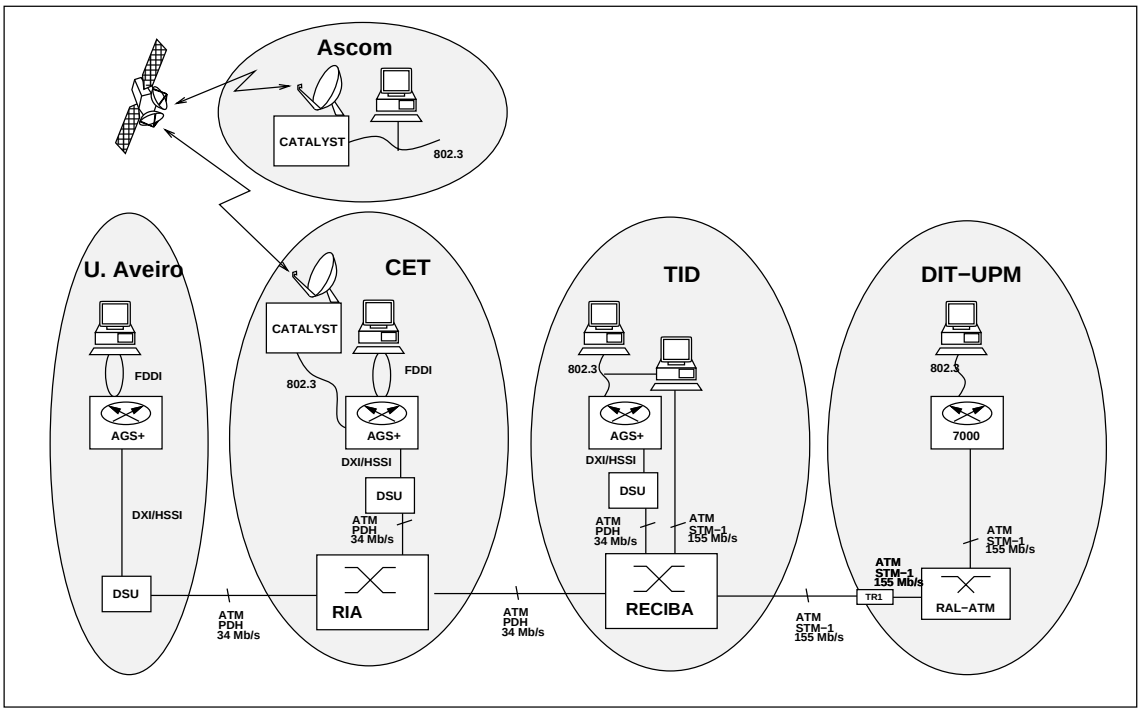

Figure 3: IP network over ATM

sulation method. Since we had two subnets, the solution to provide virtual IP over ATM was studied sepparately for each of them. Within the AAL 5 subnet (TID to DIT), as it was a virtual subnet with only two endpoints, the ATM network was treated as a point to point link. Within the AAL 3/4 subnet it was established a connectionless service following the indirect method. For this purpose, it was established a mesh of direct permanent virtual circuits between the three sites. The DSUs provided the connectionless service by relaying datagrams over the appropriate PVC within the mesh. Fast resource allocations was not possible because it was not supported by the ATM network (actually it did not support signaling at all), but the small size of the network made it manageable with manual configuration of circuits.

4.3 What Layer Multicasting?

Narrowband reliable multipoint data traffic was one of the requirements imposed by the applications. An approach to building reliable multipoint data traffic could have been to install a reliable multicast transport protocol[**?**, **?**, **?**]. However, a number of reasons led us to use a mesh of point to point connections over TCP. First, the small size of the network (five end systems) allowed a simple management of the mesh of TCP connections. Second, relative ordering of events was not a relevant issue for the application components, inasmuch as it was handled using source ordering with a token protocol. Finally, we considered that we already had too many diverse technologies in our babel-network.

Another requirement from the applications was multicast, unreliable, multimedia traffic: video and audio traffic from one source had to be sent to all the other sites. Multicasting of unreliable data is far simpler than the case for reliable data, because in the former it is not necessary to deal with the relative ordering of data from different sources. Video and audio traffic may be sent in a connectionless, unreliable, way and for this reason both were sent encapsulated within UDP datagrams to be delivered to several destinations.

Multicasting of data at the network layer is a very good solution for the multicasting of UDP datagrams because it allows the avoidance of two undesirable effects: first, it avoids loading the end

system with the burden of resending data several times; second, it avoids loading communication links with several copies of the same data. Within the network layer, better performance is obtained by making the replicas at the ATM layer (notice that ATM is being considered a network access protocol) than at the IP layer. ATM switches perform data copying in hardware, delivering copies to the output ports with minimal latency and processing overload. However, even if multicast is made at the ATM layer, it is mandatory to use multicast IP features. In particular, unless class D addresses are used, replication of datagrams at the ATM layer would result in the same unicast IP datagram being received at all destinations, and consequently discarded at all of them save for the one with matching destination IP unicast address.

The solution we envisaged was to perform multicast at the ATM layer by using point-to-multipoint ATM circuits and class D IP addresses. Unfortunately, Cisco routers did not have at that time the capability to handle multicast IP. Although Cisco provided us with beta software versions, it was judged too risky to rely the Summer School on these very preliminary versions.

To overcome the unicast IP restriction of the routers, we initially attempted to configure the routers as bridges, in order to relay frames with group MAC addresses (containing class D address datagrams), at the MAC layer, between the LAN segments to which the end systems were connected. Since the bridges would operate between LANs dedicated exclusively for the CSCW application, the flooding of group frames performed by bridges was not an issue of concern. This approach was unsuccessful because of configuration problems in the ATM interface of the routers.

Still with the same objective, we attempted quite an unusual, although potentially effective, approach: configuring the routing tables in the Cisco routers with static routes for class D addresses that would be bound to point-to-multipoint ATM circuits. This was possible because routers could forward class D addresses similarly to unicast addresses. Of course, it was necessary to use a different class D destination address from each data source, in order to be able to forward in a different way depending on the data source. One of the problems with this approach was that workstations encapsulate class D IP datagrams in MAC group address frames, while the router interfaces were not programmed to accept them. Several partially successful methods to solve the problem, such as modifying the SUN stations to encapsulate class D IP datagrams within unicast frames directed to the respective routers, were attempted, but results were not considered secure enough to be fully reliable.

Giving up the possibility of multicasting at the ATM layer, we tried to perform it at the IP layer by setting up tunnels between mrouted daemons in the SUN machines. This meant that routers would forward IP unicast datagrams in a conventional way, while mrouted daemons would replicate the datagrams in software, resending them as appropriate in encapsulated IP. This approach implied either installing mrouted at all end-systems, or, installing mrouted in a dedicated SUN station. The first alternative introduced more processing load in end-systems, which already were quite loaded. The second alternative introduced two extra routing hops for each traject, worsening the performance of the IP service. Moreover, we encountered some compatibility problems between mrouted and the Solaris SUN station acting as a router at TID (that we intended to use as a multicast router without introducing an extra hop), and also between mrouted and some FDDI drivers used in the end-systems.

The final solution was to perform multicast at the application layer, by coding a simple and efficient multicast daemon, irouted (isabel routing daemon), over UDP using static multicast routes. Irouted implements a static and hierarchical multicast network over an IP unicast communication service. Therefore, one irouted daemon should be started at each appropriate point to distribute information between all applications in the group.

Irouted avoids replication of data in communication lines. To use this approach for a particular network structure, it is required to decide the appropriate spanning tree that connects all end systems over the network. Then, irouted daemons are placed in the end systems placed at the nodes

of the spanning tree. Each irouted daemon makes copies of the UDP datagrams, forwarding them to all lines in the node of the spanning tree, save for the line from which the datagram was received. Startup files of each irouted have to be manually set, indicating the IP addresses and UDP port of the each next hop destination, either another irouted daemon or an application. Our network had two nodes where irouted was placed: TID and CET. The protocol architecture for three end systems, CET, U. Aveiro and Ascom, linked to the rest of the network is shown in Figure 4.

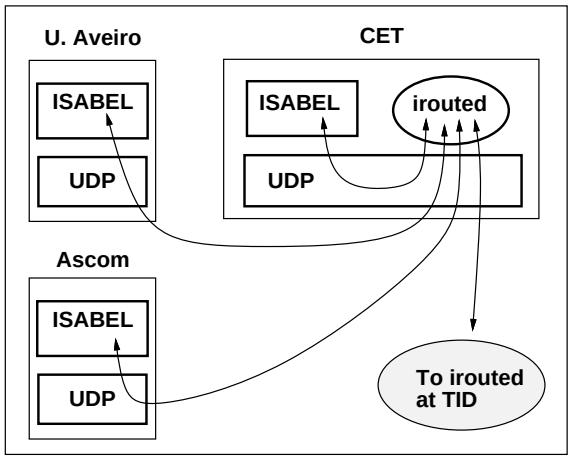

Figure 4: Architecture at three end systems

The advantages of using irouted is that static routes generate neither traffic, nor route processing overload. The obvious disadvantage for generalizing this solution is that manual reconfiguration is required to adapt the system to topology changes.

It has to be noticed that the logical structure being used to link end systems is not the same as the physical one. This causes two undesirable effect. First, several IP routes to/from one end system may be mapped over a single LAN segment connecting the end system to its router. Then, it is true that a single copy will be sent by the router over each WAN line, but, several copies are sent over the LAN segment. A second effect is that even if the end system has to send a single output copy, the LAN segment transmits the input and the output copy.

The throughput allowed by this final design was first limited by the AGS+ capacity at TID and CET, and the two Ethernet segments in the paths from TID to DIT-UPM, and from CET to Ascom. Moreover, the second effect of application layer multicasting just described caused that the Ethernet situated at TID limited the multicast network throughput to around 3.5 Mb/s. Throughput, delay and delay jitter produced acceptable video quality, but not so much for audio. The absolute delay objective (less than 200 milliseconds) was achieved between any pair of sites, save for Ascom, because of the satellite transmission delay. This was not a severe problem during the conference because panels were only held between U. Aveiro and DIT-UPM, and therefore the dialogue with the remaining sites was just *question-answer* style, which is more tolerant to delay than discussions as the ones held in panels.

5 Conclusions

Regarding ATM, it may be concluded that it provides excellent performance and capabilities for integrating different types of traffic[1]. However, ATM still has a way to go in order to be manageable in a production environment. The prototype ATM network that we set up required detailed planning and manual configuration at the ATM layer, and IP over ATM layer. Some notable improvements have been made from the time this experience was carried out and today, in particular, progress in the Available Bit Rate scheme, in the ATM-LAN, and in IP multicasting over ATM.

If ATM is to be an effective internetworking technology it is necessary that it incorporates the ease of use and flexibility characteristic of current LAN technologies. Lack of standards is not the only problem, as we also had to suffer the consequences of having too many. It has been pointed out the problems that arose from incompatible ATM physical layer interfaces between access equipment and switch ports. This is not a casual situation: the existence of a consider-

[1]Circuit emulation was also being used in the Summer School for supporting video-conferencing over commercial H.261 videocodecs, but it is not considered to be within the focus of this article

able number of ATM physical layer interfaces (e.g. STM1, STS1, TAXI, E3, combined with the choice between single-mode fiber, multi-mode fiber and UTP) can sometimes act as a barrier for the deployment of the technology. It has also been pointed out the problems caused by the existence of a number of incompatible IP encapsulation schemes. It is true that not all the encapsulation methods address exactly the same problem, and for this reason one could say that there is no duplicity, but the fact is that equipment with quite similar functionality, such as our routers, could not inter-operate because of the diversity in encapsulation methods.

Concerning multimedia traffic, several conclusions may be drawn on three different aspects: types of multimedia traffic sources, aggregated throughput required, and the need for multicasting.

On multimedia traffic types, it may be said that audio traffic requires lower bandwidth than video, but it is much more sensitive to data loss and delay jitter, even in an uncompressed form. The aforementioned sensitivity refers to the psicological perception of the received signal, this is, the audience in the Summer School did not felt uncomfortable because of occasional video frame losses or image glitches, however, minor disruption of sound quality was a severe cause of concern.

Regarding multimedia traffic volume, it should be pointed out that studying the correlation between traffic sources is important because in this type of applications the network is fed from a small number of sources with large individual bandwidth requirements, instead of from an aggregation of many sources with small individual bandwidth requirements.

In respect to multicasting, it is necessary to differentiate between reliable and unreliable multicasting service. Reliable multicasting is not so much related to multimedia as to group-ware applications, that will certainly require a reliable multicasting service. The current approaches to reliable multicast are mainly based on placing reliability at the transport layer, on top an unreliable multicast network service. The incognita still remains to be the precise ordering requirements that will be derived from the applications. Source ordering, as required by the ISABEL application, is simpler to be provided than causal ordering, which in turn is much simpler than total ordering. Independently of the ordering requirements, it is necessary that a single protocol emerge as a widely accepted standard to move the users and developers on building over it.

Concerning the unreliable multicasting service, video or voice multimedia applications do not require reliability, nor ordering requirements, and therefore can be treated with a datagram multicast as long as it provides good enough performance in terms of bandwidth, delay and delay jitter. The current solutions based on multicast IP over point-to-multipoint ATM circuits satisfy the requirements with sufficient performance.

Our final conclusion is that both broadband and multimedia technologies are advancing hand in hand at a fast pace, and in the short future will be in a state mature enough for widespread corporative deployment.

Acknowledgments

Stephen Casner from ISI-USC, Ron Frederick from Xerox PARC, Dino Farinacci from Cisco, plus all the technical team at DIT-UPM, TID and CET must be mentioned in this article for their excellent suggestions and comments that contributed to produce the final design, as well as to explore some alternative solutions.