A Generic Service Management Architecture for Multimedia Multipoint Communications

Constant GBAGUIDI, Simon ZNATY, Jean-Pierre HUBAUX
Swiss Federal Institute of Technology
Telecommunications Laboratory/ Telecommunications Services Group
CH- 1015 Lausanne- Switzerland
Phone: +41 21 693 56 13 Fax: +41 21 693 26 83
E-mail: {gbaguidi, znaty, hubaux}@tcom.epfl.ch

Topic: Service Management.

Abstract

With rapid advances in multimedia processing technologies, and in high bandwidth network technologies, new kinds of multimedia applications are now emerging. These applications require real-time, multimedia, and multipoint interactions involving multiple communicating entities and leading to a need of more sophisticated communication control functionalities. In this paper, a generic integrated and flexible service management architecture for multimedia multipoint communications is proposed, which offers three main generic functions, namely Quality of Service (QoS) negotiation, QoS monitoring and control, and session management.

1. Introduction

The availability of new technologies both in the network, where broadband transport facilities are being introduced, and at the user site, where terminals are providing for increasing intelligence and presentation capabilities, offers an unlimited number of opportunities for the introduction of new telecommunication and information networking services. Among these services are multimedia entertainment services, spanning from video on demand to interactive games, and communication services such as video conference and broadband virtual private networks.

A key issue to be tackled for providing such services is multimedia multipoint communications. For the time being these communications are not treated in an *integrated* way, that is, existing architectures do not cover the management of application resources, multimedia resources, and network resources altogether. As these resources cooperate to fulfil the users expectations, QoS requirements should not be defined statically for each given resource. For instance, when errors occur in the network resources, multimedia resources such as codecs must set their compression ratio to a low value. As a matter of fact, a low compression ratio introduces much redundancy so that the transmitted information can recover much easily from errors. The management architecture is expected to deal with this QoS adaptation between network and multimedia resources. Currently, this issue is not yet addressed.

In addition to integration, *flexibility* is another property to be required from management architectures. *Flexibility* allows for managing dynamically multimedia multipoint communications, in terms of communication session configuration (users participating in the session, involved media) and quality of service (QoS). Moreover, a *flexible* service architecture is not service-dependent and must be generic in order to be instantiated for any particular service.

To reach those objectives, namely *integration* and *flexibility*, both multimedia applications and broadband networks capabilities are to be considered. However, most of research works deal with either the former or the latter. Among these works, we can mention [2-5].

In [2], a classification of multimedia resources (also called special resources or media handling resources) is defined and a communication management architecture which exhibits two kinds of fabrics

is proposed: SWFs (Switching Fabrics), concerned with classical switching functionalities (e.g., ATM switches), and MMSFs (Multimedia Multipoint Service Fabrics) concerned with value-added multimedia functionalities. Emphasis is put on meeting the users requirements, in terms of the possibility for the users to choose multimedia resources that best suit their communications needs. Although of prime importance, QoS issues mentioned above are not considered to provide the architecture with the *integration* property.

In [3], a generic multimedia service is described, in terms of service element (SE). Each SE is a "self-contained part of the service" and is composed of operations whose temporal ordering is specific. Two kinds of SEs are identified: confirmed SEs (SEs whose execution requires resource allocation) and unconfirmed SEs. Within a communication session eight service elements are outlined: call establishment SE, add users SE, add media SE, attach media SE, detach media SE, remove media SE, remove users SE and call release SE. Although it does not put emphasis on the methods for performing those service elements (e.g., which mechanisms should be achieved for adding media), that paper presents a broad view of the users requirements on a multimedia service.

In [4], the design of a QoS broker between multimedia applications and the network is presented. However multimedia resources are not mentioned, although they have a major impact on QoS. For instance, a high compression ratio decreases the needed bandwidth, and introduces more stringent constraints upon bit error rate than a lower compression ratio does.

Moreover, the Telecommunications Information Networking Architecture Consortium (TINA-C)has defined an "open" architecture for telecommunication services in the emerging broadband, multimedia and "information super-highway" era, based on distributed computing, object orientation, and integration of Intelligent Network (IN) and Telecommunications Management Network (TMN) [5]. Nevertheless, integrated multimedia multipoint communications are still under study. Even though TINA-C mentions the need of using "information converters" in a multimedia context, a QoS management framework taking into account the impact of these resources is not yet addressed.

In this paper, emphasis is put on multimedia service aspects, spanning from multimedia resources to communication management units. Undoubtedly, it is important to study carefully the service level before addressing networking issues particularly eventual effects of multimedia communications on connection management. On the contrary of the research works mentioned above, we first identify the generic functions to be held by call control, and then propose an integrated and flexible architecture to fulfil these functions. The structure of this paper is as illustrated in Figure 1.

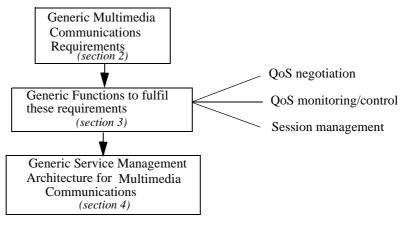


Figure 1: Structure of this document

In section 2, requirements over multimedia communications are outlined. Generic functions to be supported by the management architecture are derived from those requirements in section 3. This leads to the definition of a generic integrated and flexible service management architecture for multimedia

2. The Generic Requirements of Multimedia Multipoint Communications

Depending on the supporting service, constraints on multimedia multipoint communications are very different. For instance, telemedicine services have more severe requirements than classical videoconferencing services. However, both exhibit some common users expectations which are:

- The ability to manage dynamically the achieved QoS of each service component, may it be the audio, the video, the graphics, the picture information, and so forth. As QoS is highly subjective, it may be tuned by each party involved in the communication session. This requirement implies QoS adaptation, i.e., the possibility for end-systems to adapt the QoS to the behaviour of the network.
- The possibility for users to select some resources (network resources as well as media handling resources) that best meet their needs. As in a multimedia context, quality of service also depends upon factors such as coding techniques, quantization, and so on, the users must be proposed to choose resources according to their expectations and the capabilities of their terminal equipments.
- The possibility to interact with another user whose equipment has differing characteristics. The user expects the application he is running to deal with incompatibility problems anyhow. Hence, an MPEG video format emitter may transparently feed M-JPEG format receivers, despite the format incompatibility between the communicating parties.
- The communications network must allow any user to select the service components he wants: a user without any screen for visualizing a videoconference must be allowed to select the single audio component and interact with his partners anyway.
- The possibility for users to join in or withdraw from a communication session according to certain policies, such as agreement of all interacting partners before admitting a new user.
- The possibility to set such quality of service parameters as the synchronization tolerance between the service components, the echo sensitivity, the burst sensitivity, and so on.

The above outlined constraints require particular management schemes. The ensuing section analyzes these schemes and defines functions necessary for their achievement.

3. Generic Functions Needed within the Service Management Architecture

Service management may be defined as the intelligence, the set of mechanisms to be held in order to come up with the requirements over the provided services. Those mechanisms are inter-dependent, as they have to interact in order to fulfil the users exigencies. Therefore, mechanisms that are related can be grouped into *generic functions* that provide the service management with an organized view. As call control is a major building block that must be carefully studied when designing service architectures, we attempt herein to bring out the main functions that must be included in this important building block.

Three core *generic functions* are identified, that cover all requests to a multimedia network (this is the overall network that supports the multimedia communications). Those functions denote the meaning of the users requests to the multimedia network. They are *QoS negotiation*, *QoS monitoring and control*, and *session management*. Each of them consists of a set of mechanisms that should be achieved for performing a user request. For instance, a call establishment service element implies QoS negotiation, and then the related set of mechanisms. The following subsections focus on those mechanisms for each

defined function.

3.1 QoS Negotiation

In this paper, we represent a communication session by a collection of LCGs (*Logical Connection Graphs*). LCGs are directed graphs that describe clearly and flexibly the connectivity among multiple logical resources[9]. These are logical and graphical views of physical resources. Examples are given in section 4.

The objective is to define completely all LCGs attached to a given session. Reaching this goal means negotiating the resources (network resources and special resources) needed for establishing the communication. Special resources are of three kinds[2]:

- *Mediation resources* are devices that carry out miscellaneous functions to deal with the interfaces' requirements between different resources. Encoding/decoding devices, synchronization devices, echo suppression or cancellation devices and multiplexing/demultiplexing resources belong to this group;
- Heterogeneity handling resources are devices providing useful services to overcome heterogeneity problems such as format type incompatibility and information type (audio, video, text, etc.) differences;
- *Multipoint communication resources* are used to provide multicasting, combining (or bridging) and selecting services. These elements are responsible for the duplication of data needed, e.g., in a video-conference, since such a service may require the user streams to be duplicated before any retransmission.

Thus, the first two kinds of resources deal with multimedia issues, whilst the last category mostly covers multipoint aspects.

However, prior to QoS negotiation, the application-defined QoS should be translated into resource-defined QoS, i.e., QoS meaningful to resources, by means of a QoS translation unit.

Moreover, one LCG is defined per medium, such that, within a communication session, there are an audio LCG, a video LCG, a picture LCG and so forth. This separation of media enables a user without any video equipment to participate in a videoconference somewhat by receiving and paying for the single audio information. As a result, QoS negotiation partially solves equipment heterogeneity.

Nevertheless, selecting a resource is not so easy. Resources capabilities are to be considered. In this respect two appropriate tools, borrowed from [6] and [7], can assist in the estimation of resources capabilities versus the services these resources are able to carry out. Those tools are *schedulable region* (SR) for network resources and *multimedia capacity region* (MCR) for multimedia resources. SR and MCR respectively relate to traffic and service classes. A traffic class is characterized by some statistical properties and QoS requirements (e.g., processing duration and error rate). For instance, the most stringent traffic class is characterized by 0% as error rate, and no processing duration so that packets received at the resource interfaces are not subject to queuing. A service class is defined in the same terms for multimedia resources [6]. The schedulable region of a resource is "the set of points in the space of possible loads for which a given quality of service is guaranteed"[6]. The possible loads of interest for a network are traffic classes. The MCR is defined similarly to the SR with service classes as loads.

Schedulable region and multimedia capacity region are powerful tools, but the characterization of service classes, mainly based on encoding types, gives rise to many classes and does not facilitate the determination of the regions of interest. Even though the number of encoding types is currently huge, schedulable region and multimedia capacity region concepts could give a needful logical view of physical devices, which is helpful for performing call admission control for both networking and multime-

dia resources. The main open issue is the number of classes to be considered.

3.2 QoS Monitoring/Control

QoS monitoring/control is concerned with holding the achieved QoS up to the negotiated QoS as to guarantee the expected QoS. This can be done in two ways: *QoS compensation* and *distributed resource management*.

The purpose of *QoS compensation* is to find the way(s) QoS can be shared out among the network resources and those concerned with media handling. Indeed, the performance of the latter resources can help release the QoS requirements over the network. For example, using M-JPEG for video requires higher bandwidth than MPEG encoding technique does. The intelligence to be provided by QoS compensation then consists in having an integrated view of multimedia communications that encompasses network and multimedia technologies, so as to deal with the networking consequences after altering some multimedia resources parameters, and vice versa.

The distributed resource management function is essentially a fault management function. It may be performed using the L-E (Legislator-Executor) model [8] (Figure 2) for a collection of resources or each resource alone. The legislator, after consulting the resource capacity (which in fact is the schedulable region or the multimedia capacity region) and the request intensities data store, sets rules (control policy) on accessing the resource. The executor regulates this access conforming to the control policy. It also controls the resource state and, like a watch-dog, makes accurate decisions (e.g., failure notification) about using the resource. On top of the model control parameters are set to manage the control task.

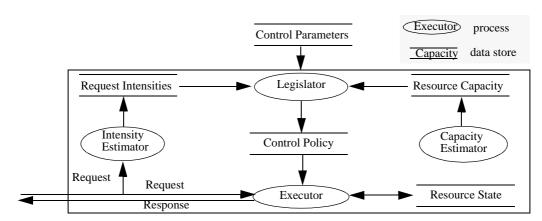


Figure 2: Resource control subsystem: the L-E model [8]

In short, QoS monitoring/control is performed in two manners: one, distributed, is the resource management function and the other, more centralized, is QoS compensation. This latter is more complex and may introduce undesirable transient effects when performed at runtime with a huge amount of resources. Transient effects are due to the fact that the set of resources involved in the communication has changed owing to a resource that failed in providing its QoS part. In the opposite of QoS compensation, the distributed feature of the resource management function makes it more robust against transients effects. Nevertheless, the scope of that distributed resource management is limited to "weak" failures. In cases where a resource cannot recover from its failure and cannot be replaced by any similar equipment, the only function available is QoS compensation which then initiates a new QoS translation.

3.3 Session Management

Session management provides mechanisms that enable the user to change the configuration and attributes of an on-going communication session. Among operations for changing those attributes and configuration are medium/user admission control, medium/user removal from a session and QoS modification. Most service elements described in [3] belong to this category of operations. Using medium-based LCGs, i.e, one LCG per medium, facilitates the achievement of management operations, since the session management function results in modifying existing LCGs (in case of admission or removal of a new user) or building new LCGs (in case of admission of a new medium). Admission control becomes a problem when the communication session could not admit a new user or medium for whom there is no more resource available, unless the users participating in the session agree with decreasing their current QoS. Therefore, admission control mainly consists in modifying existing LCGs.

The generic functions described above are not just co-existing but they also need to co-operate. Thus, session management involves QoS re-negotiation and QoS monitoring/control. Yet, QoS negotiation presents a static aspect, as it consists in negotiating a required *constant* QoS, whilst QoS monitoring and control rather presents a dynamic feature as it keeps an existing configuration, especially the achieved QoS, by adjusting the resources efficiency. Session management presents a much more dynamic feature, since the QoS is dynamically changed as well as the communication session configuration. An open issue is to be studied is how to modify the achieved QoS without damaging the ongoing communication.

As previously stated, the generic functions are necessary for meeting the users requirements. In the next section, an architecture is proposed with regard to these functions.

4. A Generic Service Management Architecture for Multimedia Communications

The proposed architecture (Figure 3) enables the realization of the three generic functions described in section 3, namely QoS negotiation, QoS monitoring/control and session management. After specifying the roles taken by the architecture components, the way each function is carried out by the architecture is described, and the proof of the latter's flexibility brought up.

In order to allow for the sharing of multimedia resources among several users, a Customer Premises Equipment (CPE) is introduced between end-systems and public networks. Therefore users do not need to buy costly equipments before accessing multimedia services.

4.1 Architectural Units

Before addressing the overall functionality (provision of the generic functions) of the architecture components, we first describe these.

The user agent represents the profile of the service user, in terms of the latter's identity or some other features like the user's visual acuity, and so forth. Those features may be needful for making QoS user-specific, thus expressing the subjective nature of QoS.

The application provides the user with an interface for accessing the service.

The user session management unit is concerned with the management of the user session at the end-system; it is also in charge of selecting the terminal media handling resources needed to fulfil the QoS requirements. Moreover, the described unit interacts with the service session management instance for creating and managing the communication session.

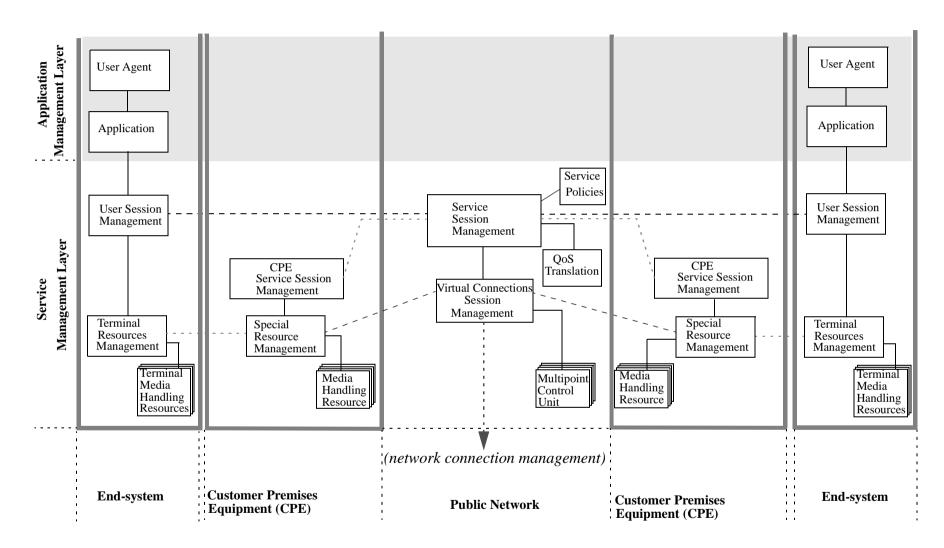


Figure 3: Generic Service Management Architecture for multimedia communications

The terminal media handling resources are the multimedia equipments embedded by the end-system.

The CPE service session management unit manages the session part within the CPE.

The CPE special resources management unit is responsible for selecting and managing the special resources within the CPE and for curing resource failures somehow.

The Terminal Resources management unit manages the special resources within the terminal.

The virtual connection session management unit manages the media LCGs used for the purpose of the communication. This management task necessitates interactions with the network and the CPE special resources management units to be involved.

The service session management unit manages the whole session, interacts with the network and the service policies to establish the bill, for instance. Another important decision it takes in concert with the service policies is how to select the networks to be used for the communication with the objective of reducing the communication cost. Service policies may serve to feature commitments between the service provider and some network operators. The service session management unit is also responsible for creating the environment needed for managing the communication sessions.

The functionality of the *QoS Translation* unit is to translate application-required QoS into resource-defined QoS and eventually propose many issues for selecting the resources to involve in the communication sessions.

The multipoint control unit (MCU) is well described in [10] for bandwidths up to 2Mb/s, but this bandwidth restriction takes nothing off from the functionality. However, H series recommendations (e.g., [11]) do not consider sharing of special resources between many terminals and distinguish between primary terminals, whose capabilities match those of the used MCU, and secondary terminals whose capabilities are far different from those of the used MCU. For the sake of conciseness, the MCU is described as containing video combiners, audio mixers, multiplexing/demultiplexing resources, network interfaces, audio/video switching resources, echo handling unit, among the most important components. MCUs are discriminated by their processed audio/video features, their bandwidth, the number of simultaneous conferences they could handle, the number of participants in each conference, and so forth.

4.2 Realizing the Generic Functions

4.2.1 Realizing QoS negotiation

Considering the very first QoS negotiation that takes place during the communication establishment, a user agent creates a user session by running its multimedia application (Figure 3). The created session submits the desired LCG (Figure 4a) to the service session management instance. The latter is responsible for controlling the overall communication. It first translates the application-required QoS to resources-defined QoS using a QoS translation unit. A resource-defined QoS is a QoS expression made meaningful to the resources to be involved in the processing of the communication. The example illustrated below assumes that the session initiator negotiates for a video mixing. Thus, users are allowed to choose resources that best meet their needs. Note that initially the session initiator knows nothing about the capabilities of the terminals owned by its potential partners (this is depicted by the question marks in some end vertices in Figures 4a and 4b).

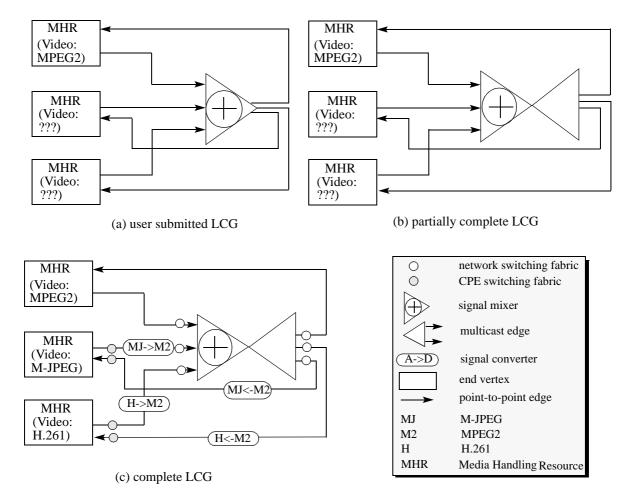


Figure 4: Evolution of a video LCG towards completeness

Once all resource-defined QoSs are calculated (thanks to the QoS translation unit) the *Virtual Connection Session Management* (VCSM) instance starts negotiating for the resources needed by interacting with the network and the *CPE special resources management* (SRM) instances. The VCSM is responsible of involving the networks selected according to the *service policies*. These also include billing and security policies to be applied to the communication messages. If any multipoint control unit (MCU) is needed, it is considered by the VCSM instance (Figure 4b, the signal mixer and the multicasting functionalities may be provided by a MCU). Networks stemmed from the *service policies* provide connectivity to support the multimedia communication. This connectivity adds information to the LCG (switching fabrics vertices).

However, the medium LCG is not yet complete at this stage. It still needs to check the capabilities of all partners in order to be managed by the VCSM. It is up to the latter to interact with the SRM instances within CPEs concerned with the communication session to select the special resources to use further on (Figure 4c). Those instances select the suitable resources to cope with the supplied resource-defined QoS using L-E models of resources (subsection 3.2). Before any selection the SRM instance first checks the capabilities of the terminal to connect. If these are not sufficient to support the communication, then the SRM instance chooses among the resources it manages the adequate ones for processing the information.

Together, operations performed by the network, SRM and VCSM instances lead to a complete medium LCG. It is worth noting that the proposed framework makes no assumptions about the way networks to be involved are selected (that is up to the *service policies* unit). Moreover, it does not treat any particu-

lar service. That proves the *flexibility* of the architecture in Figure 3.

Throughout the above example, the strength of medium LCGs has been well stated. They permit to deal properly with media handling within a communication session and adds flexibility to the architecture.

4.2.2 Realizing QoS Monitoring and Control

QoS monitoring and control is concerned with keeping the achieved QoS close to the negotiated QoS. Monitoring is different from management in that its role is not to change the QoS, but to guarantee the quality of the communication. Hereafter is shown the way the two means (QoS compensation and distributed resource management) for performing the monitoring/control function are realized.

The QoS translation instance deals with the compensation function by suggesting different schemes to cater for the QoS requirements. For instance, when an encoding technique becomes unsatisfactory, due for example to network errors or low processing time at end-systems, another encoding type must be suggested. The compensation function is rather complex. The other method for QoS monitoring, namely the distributed resource management function, is simpler and should be first used before any compensation mechanism.

The resource management function is essentially a fault management function. In the proposed architecture, it is performed in end-systems (by *Terminal Resources Management* -TRM- units), in CPEs (by SRM) and in public networks (by network connection managers), i.e., every site whose resources are involved in the communication. The management units for network resources are not visible from application and service points of view. Therefore, they do not appear in the service management architecture. In the case of failure of the terminal resources, the TRM unit must inform the related SRM in order to find among special resources within the CPE those suitable for a replacement. In general, after any failure or recovery, the concerned medium LCG should be changed accordingly.

4.2.3 Realizing Session Management

Session management is mainly concerned with user/medium admission control, user/medium removal and QoS modification. Removal operations are generally not very difficult to perform in the opposite of admission or QoS modification operations. These have the particularity of modifying the media LCGs a great deal by implying QoS re-negotiation, and thereby re-negotiation of the resources to be involved in the communication session.

Medium admission requires the design of a brand-new medium LCG. On the other hand, user admission is usually expressed by the addition of new end-vertices to existing LCGs.

The main problem to be solved when managing a session is how to negotiate QoS in order to admit a new instance, medium or user as well. Two cases may occur: in the first, there are enough resources so that there is no problem, whilst in the second the lack of resources leads to implement a mechanism that may run successfully or not. In the latter situation, the participants in the communication session should be asked for agreement not only with decreasing their QoS but also with the acceptable decrease amount. If this is acquired, a suggested mechanism is to find out the remaining capabilities of each involved resource (after decreasing) in order to "spread out" the new user's QoS over these resources. A solution from the proposed architecture, as described in a previous subsection, is for the QoS translation unit to suggest many different ways permitting to meet the QoS requirements. Thus, as a consequence of bandwidth decrease, a sufficiently high compression ratio can be recommended. The QoS translation unit should propose a suitable compression ratio. The way this is implemented is a matter of further research. However, the entities and concepts (mainly LCG, schedulable region, L-E model, multimedia capacity region) to use are visible in the generic functions.

5. Conclusion

The major achievements of this work are: identification of three generic functions that must be supported by any multimedia service management architecture, proposal of mechanisms for each generic function, and proposal of an *integrated* and *flexible* service architecture for multimedia communications featuring all proposed mechanisms. Although this architecture covers all users requirements, there still are many open issues not yet addressed. First, the QoS translation unit needs to be studied carefully, because it is the basic building block on which the architecture relies, especially when dealing with multimedia applications.

Another important issue is to define an information model of special resources in order for these to be manageable. The design of such a model requires the study of characteristics of those resources from architecture and management points of view. Moreover, synchronization issues need to be clarified. Synchronization is a relationship between one or more media, or between interacting parties. There are three kinds of synchronization: intra-medium synchronization (between information units pertaining to the same medium), inter-media synchronization (between two or more media, e.g., audio and video synchronization) and inter-party synchronization (between interacting parties). The latter kind of synchronization is intended to preserve the sequence of events occurring at the communicating parties.

Finally it is important to study how to hold an on-going multimedia communication while negotiating a new QoS. In other words, how to deal with transient effects? Once these issues are solved, the proposed architecture will be implemented in the context of the OAMS (Open management Architecture for Multimedia Services over ATM) project [14].

6. References

- [1] Connection Management for an ATM Network L. CRUTCHER, A. GILL WATERS, IEEE Network, November 1992.
- [2] *Towards Flexible Communications Management for Multimedia Multipoint Communications -* Y. KATSUBE, IECE Transactions on Communications, Vol. E77-B No. 11, Nov. 1994
- [3] Communications Systems Supporting Multimedia Multi-user Applications G.J. HEIJENK, X. HOU, I.G. NIEMEGEERS, IEEE Network, Jan./Feb. 1994.
- [4] The QoS Broker K. NAHRSTADT, J.M. SMITH, IEEE Multimedia, Spring 1995.
- [5] Building New Services on TINA-C Management Architecture J. PAVLON, L. RICHTER, M. WAKANO. 3rd International Conference on Intelligence in Networks, Bordeaux France, 1994, pp259-263.
- [6] *Real-Time Scheduling with Quality of Service Constraints* J. M. HYMAN, A. A. LAZAR, G. PACIFICI IEEE Journal on Selected Areas in Communications, vol. 9, Sept. 1991- Also available via: ftp.ctr.columbia.edu/CTR-Research/comet/public/papers.
- [7] Multimedia Networking Abstractions with Quality of Service Guarantees A. A. LAZAR, L.H. NGOH, A. SAHAI- Available via ftp.ctr.columbia.edu/CTR-Research/comet/public/papers.
- [8] An Architecture for Performance Management of Multimedia Networks G. PACIFICI, R. STA-DLER. Available via ftp.ctr.columbia.edu/CTR-Research/comet/public/papers.
- [9] A Connection Manager for Flexible Specification and Transparent Location of Special Resources S.L. MOYER & D.S. ROUSE, GLOBECOM'93, pp. 1526-1530.

- [10] Multipoint Control Units for Audiovisual Systems using Digital Channels up to 2 Mbit/s ITU-T Recommendation H.231.
- [11] System for establishing Communication between Audiovisual Terminals using Digital Channels up to 2 Mbit/s- ITU-T Recommendation H.242.
- [12] Multipoint Audio and Video Control For Packet-Based Multimedia Conferencing F. GONG, ACM '94, San Francisco, CA, USA.
- [13] *EXPANSE Software for Distributed Call and Connection Control* H. BUSSEY et al., International Journal of Communication Systems, vol. 7, 146-160, 1994 by John Wiley & Sons, Ltd.
- [14] *OAMS: An Open Management Architecture for Multimedia Services over ATM* S. ZNATY, C. GBAGUIDI, J-P. HUBAUX, Internal Report, 1995.