Quality of Service Management and Control of Multimedia Applications: a Scenario and Two Simple Strategies

R. Bolla, F. Davoli, M. Marchese

Department of Communications, Computer and Systems Science (DIST)
University of Genoa
Via Opera Pia 13, I-16145 Genoa, Italy

Abstract

Management and control of multimedia traffic is considered within a scenario characterized by the presence of different network platforms, exhibiting possibly very different features in terms of Quality of Service (QoS). In this framework, the needs for control strategies, acting in the user plane at protocol layers above the transport, is examined, and their goals, possible actions, information bases and location (centralized/decentralized) are discussed. Two specific strategies for QoS control of a multimedia connection involving real time traffic (voice and video) are then introduced; one is based on a backoff procedure at the transmitter, while the other acts upon feedback from the receiver to change some features of the coding algorithm. The joint operation of both control strategies is examined in an experimental environment based on Ethernet networks, interconnected by IP routers.

1. Introduction

Networked Multimedia Applications (NMA) are characterized by the presence of a mix of different traffic streams, some of which (e.g., video) can be very bandwidth-demanding. This may not be a problem in the presence of high speed networks, especially if the network is capable of offering Quality of Service (QoS) guarantees, as will be the case with the future B-ISDN in the wide area. However, in general, several situations exist where multimedia information transfer takes place over less endowed network environments, and even over networks that only offer "best effort" transfer and delivery (e.g., over the Internet). By considering, for instance, only the local environment, a fairly general scenario today can be constituted by *shared medium* networks at relatively "low" speed (Ethernet), "medium" speed (FDDI, FastEthernet), as well as by high speed *switched* networks (ATM LAN) [1]. Capacity limitations are very likely to become a bottleneck in the shared medium subnetworks, severely limiting the number of simultaneous multimedia connections that it is possible to support. Moreover, such shared environments are today so widely deployed that they are likely to remain for some time to come, before being totally replaced by the high capacity ATM switches (which, in any case, also do not provide unlimited resources). On the basis of these considerations, we believe that network management and control tools should be widely applied, in order to guarantee Quality of Service (QoS) requirements to the maximum possible number of applications in a highly heterogeneous network environment. Moreover, as the underlying networks may exhibit very different characteristics and resource sharing mechanisms, controls are necessary that will act somehow "above" the network layer (typically, at the transport layer and higher [see, e.g., [2]).

In this paper, we first attempt to describe in some more detail a highly variable scenario, by considering the different network capabilities, and to define possible control goals and actions, different QoS perceptions, and possible control architectures. Then, two specific strategies for QoS control of a multimedia connection involving real time traffic (voice and video) are introduced; one is based on a backoff procedure at the transmitter, while the other acts upon feedback from the receiver to change some features of the coding algorithm. The joint operation of both control strategies is examined in an experimental environment based on Ethernet networks, interconnected by IP routers.

The paper is organized as follows. The next Section is devoted to the definition of our working scenario and of the general control approach. Section 3 describes the two simple strategies that we have started to evaluate in a small experimental environment, whereas Section 4 discusses some experimental results and remarks some problems. Section 5 contains the conclusions and suggestions for further research.

2. An organizational model

As we mentioned, the network transfer mode and offered services may have a great influence on the higher layer control structures that we want to investigate (in a similar way as, for instance, different grades of service, as provided, e.g., by datagram and virtual call networks, affect the complexity of the transport protocol). In order to distinguish these characteristics, we consider two different environments, which we will call "*structured*" and "*unstructured*", respectively. To the first one belong

0742-1303/95 $4.00 © 1995 IEEE

networks which already exert control actions in order to guarantee QoS requirements (in other words, we may say that these networks have non void lower layers in the management and control planes). Circuit switched networks, hybrid-TDM networks (e.g., DQDB, as regards the pre-arbitrated traffic) and ATM networks fall in this category. Since in these cases applications "sit" on a transfer mechanism where the resource assignment is already controlled, controls at the higher layers are relatively easier to perform. On the other hand, we may say the resource assignment in shared medium networks of the Ethernet type and in the asynchronous parts of FDDI and DQDB, as well as in best effort packet networks like the Internet, is "unstructured", since no QoS guarantee is ensured in these cases, and sometimes even no distinction can be made at the network access level in terms of different traffic requirements.

With this consideration in mind, we touch some points that pertain to our control structure in the following.

QoS translation

QoS is generally expressed in different terms at different layers, and "translation" of QoS requirements may be necessary, for instance, at the boundary of a structured network. In the ATM case, QoS can be guaranteed in terms of cell transfer delay and cell loss rate; on the other hand, at higher layers, QoS may be expressed in terms closer to the user perception, especially as regards video and audio quality. We refer to this as *Perceived* QoS (PQoS) [3], of which we may accept looser limits, related to the user's satisfaction. For the time being, we do not give a precise definition of PQoS objects and of their characteristics; however, we can keep in mind that higher level control goals will be based on it, rather than on (possibly more "strict") QoS requirements.

Management and Control

We adopt a distinction of management and (real time) control, based mainly on a difference in time scales [4]. In this respect, a hierarchical multilevel dynamic control structure, whereby the management plane plays the role of a higher control level (or "coordinator") and several dynamic control units independently act at the transport level, lends itself as a rather "natural" decomposition of a very complex overall control task. In this type of architecture, however, several points still need to be defined, regarding possible control actions, goals, constraints, information available for control purposes, information and computation dissemination.

Positioning of the control levels within the protocol architecture

This item has to be investigated, and the architecture of the control system may be changed accordingly. For instance, by maintaining a two-level structure, we can situate the lower control level at the transport layer, and the higher level at the session layer (or the corresponding structure above the transport layer, depending on the protocol architecture); in any case, we may suppose a "multimedia connection" to take place at some layer above transport. In a different setting, controls related to the application may be performed at a connection level, and be coordinated by a management station. The overall management of this object can actually be thought of as performed in turn by a multilevel structure, where "local" controllers (taking care, e.g., of a whole local area environment) are coordinated by a centralized decisional agent, which has a view of the connections, spanning possibly several "local" domains, as a whole.

Controlled objects and control actions

The possible control actions should be identified, as well as the control objects and goals. This is one of the preliminary research activities. Wherever a connection-oriented service operates, connection admission control is allowed. Thus, at the higher level of our control architecture, which we suppose to be situated at a protocol layer characterized by a connection-oriented service, one of the possible actions is admission/refusal of a connection request. Another possible action, which, however, deserves further thought, may consist in the (dynamic) allocation of resources (buffers, CPU, bandwidth) to certain "classes" of services (this, in turn, would imply that connection admission control takes place on a per-class basis, and that each class might be dedicated an independent admission controller). At the lower control levels (e.g., those situated at the transport layer and interacting with the transport protocol), control actions may take on a simpler form, depending on the amount of information that is available. In a connectionless transport service, like UDP, it seems reasonable to think of *decentralized* controls, based on information on the "local" domain, which is available at the transport stations individually, like, e.g., overcoming thresholds on buffer contents. This may trigger the application of backoff strategies, which might consist in some form of rate reduction (within the limits allowed by the service contract or not) achieved, as regards real time traffic, by changing some parameters in the coding algorithm or the coding algorithm itself.

Information available for control

The main reasons to set up a hierarchical control structure in this setting lies in the difference in information that may be available at different levels, mainly caused by computation and transmission delays [5]. Information dissemination may be limited by the lack of monitoring capabilities, even in broadcast environments. For instance, the capabilities of a network monitor are generally not available at common low-cost Ethernet interfaces. This leads to an inherently informationally decentralized setting (this is true even at the MAC layer, where Binary Exponential Backoff is solely based on the *individual* involvement in repeated collisions [6]). On the other hand, if a connection oriented multimedia service is performed at some point in the architecture, it becomes relatively easy to keep track of the number of connections and, in case of resource assignment on the part of some local management agent, even of their parameters and status, which allows a (partial) centralization of this particular kind of information.

"Parallel" controls (synchronization issues)

The types of control that have been mentioned so far attempt to achieve QoS requirements by means of "non-

specifically addressed" (i.e., not applying to a particular parameter or characteristics) actions. Synchronization issues in multimedia communication, either serial (consisting basically in delay jitter reduction) or parallel (media synchronization), deserve further attention and perhaps *specific* controls, even though they are indirectly affected by the others (like congestion control). This is by now a well addressed topic in the literature [7, 8], and perhaps the point that should be investigated regards the interaction of this type of control with the others within the whole architecture.

3. A simple general structure of distributed control

Some additional considerations can be done regarding the definition of QoS and of service itself. A user, at the time of connection set-up, should choose among the defined services, for example videotelephony, teleconference, teleteaching, video monitoring, audio conference and so on, and specify some desired parameters affecting PQoS. On the other hand, each device (for example workstation) has to define an audio/video multimedia service in terms of technical characteristics and QoS parameters. Concerning video, these characteristics could be the size of the video window, colour depth, frame rate, type of compression, compression quality parameters; as far as audio is concerned: sample rate, type of compression and so on [3]. QoS parameters could be the packet loss and delayed packets rates (in a packet network as considered in our experimental environment). The set of the possible values of the different characteristics can be represented by a multi-dimensional space, where each point represents a "configuration". So, the service required by a user has to be "translated" into a suitable configuration, taking into account that individual workstations may support only a sub-set of all the possible configurations, because of different hardware devices and computational power. Moreover, more than one station is involved during a multimedia communication, at least two in a point-to-point connection, so the sub-set of characteristics, where a configuration has to be found out, shall be built as the intersection of the sub-sets of all involved stations.

Each different configuration generates a different load offered to the network. Depending on the network type and condition, the resulting load should affect the QoS (packet loss and delay rate) and, as a consequence, the quality of the service seen by user (PQoS in the previous Section).

The basic idea of the distributed control scheme proposed in the following is the managing of the offered load, expressed in terms of bit-rate, and its balancing with the measured received load (the bit-rate, at the receiver, of the packets whose delay and loss rate are below fixed thresholds). The control mechanism is not limited to the set-up phase, but is applied during the entire life of a connection to follow the possible variation of the load in the network. Two different types of control have been proposed: a local one and an end-to-end one. The local control performs local measures (in our case the packet transmission delay) and uses them to change the original bit-rate. This type of control is effective only for communication on the same LAN, because only in this case measures are significant for the whole source-destination path. The second type of control operates on an end-to-end basis, by using feedback from the receiver, and it can be operative in a heterogeneous environment composed by different interconnected LANs. Both have to change the offered load, and this change is performed by choosing a new service configuration, which generates the requested bit-rate. The choice of a new configuration shall be based on the concept of minimum distance in a multi-dimensional space. The algorithm proposed in the following is a first simple example of possible solution; further studies are in progress to obtain a more structured solution as in [9], in a different context.

The scenario introduced above has been applied to develop and implement an experimental multimedia application. This application and the used control mechanism have to be considered as a first result to have feedback of the user perception; further and accurate studies are needed to improve the dynamic control mechanism.

The structure of the specific application, which has been developed for Apple Macintosh computers, will be described in the next Section; in the following, only the implementation of the control scheme is described. It has to be observed that the proposed control scheme is intended to operate above the Transport layer; in our application, the Internet Protocol Suite is used and, in detail, the UDP protocol is employed at the Transport layer for voice and video. Background data traffic (like the one generated by telnet or ftp connections) is, in our environment, uncontrollable, and it plays the role of a "noise".

Compression Type	Video Window Size	Video Quality	Frame Rate (frames/s)	Bit-rate (Kbits/s)
APPLE VIDEO 16 bit	128x96	MIN	12	425
		NORMAL	11	515
		MAX	9	550
	160x120	MIN	8	450
		NORMAL	8	570
		MAX	6	550
	320x240	MIN	3	300
		NORMAL	3	600
		MAX	2	660
JPEG 8 bit	128x96	MIN	8	70
		NORMAL	7	140
		MAX	6	315
	160x120	MIN	5	55
		NORMAL	5	135
		MAX	4	300
	320x240	MIN	2	60
		NORMAL	2	150
		MAX	2	460
JPEG 24 bit	128x96	MIN	7	85
		NORMAL	6	150
		MAX	4	250
	160x120	MIN	5	80
		NORMAL	5	170
		MAX	3	270
	320x240	MIN	2	75
		NORMAL	2	180
		MAX	1	300

Fig. 1. Maximum frame rate and the resulting bit-rate for every possible configuration.

As said above, the control system is based on the definition of possible transmission configurations, which imply different performance of the transmitter-receiver pair to follow possible load variations in the network. In this environment, a configuration has been defined by three parameters: the video window size, the type of compression and the video quality ; a fourth parameter could be included, even if not strictly independent of the others: the frame rate. The window size can assume three values: Small (128*96 pixels), Medium (160*120) , Large (320*240). The choice of the compressor is limited to three algorithms: APPLEVIDEO, JPEG-8bit, JPEG-24bit. The parameter "video quality" controls the image quality by acting on the compressor; it can assume three values: Minimum, Normal, and Maximum, corresponding to a decreasing compression ratio. The maximum frame rate depends on the compression algorithm, the video quality and the window size, but the maximum value can be decreased if the network load is out of the possible range; from this point of view, it can be considered a configuration parameter. The described situation is depicted in Fig. 1, where every possible configuration is shown with the maximum frame rate and the resulting bit-rate (Kbits/sec).

The initial configuration can be chosen by the user. This basic configuration is transmitted to the receiver that, if ready, sends a confirmation message.

After data transmission is in progress, two control mechanisms are available to guarantee Quality of Service, namely, the local control and the end-to-end control. As already mentioned, both control schemes have to choose a new configuration generating the requested bit-rate. The following variables can be defined: let x_{curr} be the current bit-rate, x_{new} the new requested bit-rate, x_{log} a logical variable used to have a trace of the requested bit-rate and initialized at the value x_{curr}. The choice of the new configuration is so performed:

1) Let $y=\delta \cdot x_{new}$, with $\delta<1$, be a threshold whose meaning will be clear in the following (in our implementation, we have used $\delta=0.05$);
2) If $|x_{log}- x_{new}| \leq y$ (no other configuration closer to x_{new} than x_{log} can be found), then the variable x_{log} is set to the value x_{new} and the algorithm is stopped without changing configuration. Otherwise,
3) the set of all the configurations obtained by varying only the compressor type, and with average bit-rate within the interval $[\max\{x_{new}-y; 0\}, x_{new}+y]$ is considered. If the set is not empty, go to the step 8). Otherwise,
4) the set of all the configurations obtained by varying the compressor type and the video quality, and with average bit-rate within the interval $[\max\{x_{new}-y; 0\}, x_{new}+y]$ is considered. If the set is not empty, go to the step 8). Otherwise,
5) the set of all the configurations obtained by varying the compressor type, the video quality and the frame rate, and with average bit-rate within the interval $[\max\{x_{new}-y; 0\}, x_{new}+y]$, is considered. If the set is not empty, go to step 8). Otherwise,
6) the set of all the feasible configurations with average bit-rate within the interval $[\max\{x_{new}-y; 0\}, x_{new}+y]$ is considered. If the set is not empty, go to the step 8). Otherwise,
7) the algorithm is started again from 2), after increasing y of $\delta \cdot x_{new}$.
8) the configuration with the bit-rate closer to x_{new} is chosen; the variable x_{curr} is set to the chosen bit/rate value and the variable x_{log} is set to the value x_{curr}.

The main aspect of the algorithm introduced above is the introduction of a decreasing priority order among the configuration characteristics; this order is: type of compressor, video quality, frame rate and window size. The use of a logical value of the bit-rate instead of the actual value is introduced to avoid possible deadlocks in the algorithm which have been matched during the test phase.

Concerning the end-to-end control, at the receiver the input bit-rate is compared with the output bit-rate at the transmitter (contained in each packet). Both the transmitter and the receiver bit-rate are averaged over a fixed interval to avoid transient situations. If the difference between the two bit-rates is larger than a certain value, an intervention is needed, because the transmitted data is lost or excessively delayed. The remote receiver communicates to the transmitter an alarm signal containing the receiver input rate, which is the new requested bit-rate x_{new}, and a new configuration with the bit-rate x_{new} is looked for, by using the algorithm introduced above.

Compression Type	Video Window Size	Video Quality	Average Frame Transmission Delay (msec)	Standard Deviation
APPLE VIDEO 16 bit	128x96	MIN	11.442	0.301
		NORMAL	14.258	0.346
		MAX	14.474	0.242
	160x120	MIN	12.780	0.309
		NORMAL	16.159	0.209
		MAX	15.667	0.462
	320x240	MIN	16.176	0.529
		NORMAL	18.718	0.766
		MAX	19.282	0.291
JPEG 8 bit	128x96	MIN	10.069	0.527
		NORMAL	12.004	1.209
		MAX	14.625	0.775
	160x120	MIN	10.386	0.171
		NORMAL	13.346	0.922
		MAX	16.604	0.557
	320x240	MIN	16.381	0.596
		NORMAL	20.446	0.980
		MAX	19.157	0.405
JPEG 24 bit	128x96	MIN	10.444	0.328
		NORMAL	11.673	0.337
		MAX	16.486	0.328
	160x120	MIN	10.767	0.564
		NORMAL	12.970	0.633
		MAX	17.665	0.672
	320x240	MIN	16.590	1.070
		NORMAL	19.598	0.588
		MAX	21.335	0.231

Fig. 2. Average "reference" value of AFTD (R-AFTD) and its standard deviation (SD-AFTD) for configurations with maximum frame rate.

As far as the local control is concerned, the control action is performed by the transmitter by using a local measure of the average frame transmission delay (AFTD) and applying a backoff-like algorithm. This value depends on the load conditions, so a "reference" value has been measured in the complete absence of background network traffic. The average "reference" value of AFTD (R-AFTD) and its standard deviation (SD-AFTD) are depicted in Fig. 2 for configurations with maximum frame rate.

The current AFTD value is compared with the reference value; when the current AFTD is out the interval R-AFTD±SD-AFTD, a new configuration with a bit-rate x_{new} equal to 90% of the logical value (x_{log}) is looked for, by using the same algorithm as in the end-to-end case. After a period of time where the above control is not requested to act, the output bit-rate is increased, by looking for a new configuration with a bit-rate x_{new} equal to 110% of the logical value.

The next Section is dedicated to a description of the multimedia application and to the presentation of the first preliminary results obtained.

4. Application description and results

The application has been developed for Apple Macintosh computers, and uses the Apple Operating System (AOS) version 7.1, and the Apple multimedia AOS extension (QuickTime) to manage all local multimedia operations (image acquisition, local visualization and compression/decompression). The computers used for the experimental set-up are two Macintosh Quadra 900 without any special hardware device except for a RasterOps 364TV frame grabber board, which is able to acquire a maximum 30 images/s, with a dimension of 640x480 pixels and 24 bits of colour depth. All other operations, (compression, for example) are performed by software.

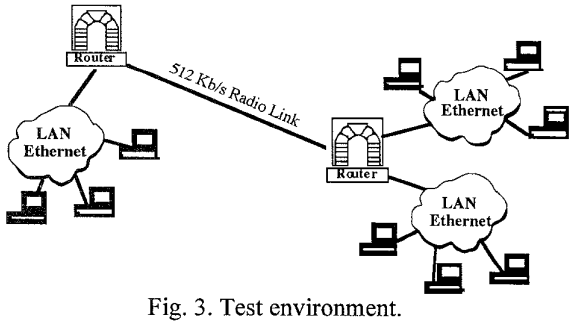

Fig. 3. Test environment.

The test environment is shown in Fig. 3. It is composed by two local Ethernet LANs connected to a router. This router is connected, via a dedicated synchronous radio channel at 512 Kbits/s, to a remote site (50 km apart), where there is a second router connected to another Ethernet LAN. By using this environment, three types of tests have been performed: a test on the LAN only, a test over a connection between the two LANs connected by a router, and a third one between the two remote sites. The different tests are performed by letting the load change in the network, as in normal working situations.

The first results obtained are related with the impression that the control strategies have made over the users. Concerning the tests over a LAN, it has been observed that, in normal conditions, the network load alternates periods of ordinary load, and overloaded short periods (during ftp, for example) where no multimedia communication is really possible. In this last situation, both local and end-to-end control are not really effective, while controls are satisfactory, from the user point of view, in ordinary situations. The behaviour in the two LANs and remote-sites communication has not been tested enough to draw some conclusions, but it has been already noted that the local control is not very effective in these environments, while the end-to-end scheme is surely better suited for these situations.

While further experimental trials are being performed in this setting, a centralized management scheme has been implemented, based on SNMP-2, by adding specific multimedia related objects in the Management Information Base (MIB). This is currently being interfaced with the decentralized controls, in order to obtain a coordinated multilevel control hierarchy.

5. Conclusions

In this paper, a scenario has been proposed within a multimedia environment. Topics as QoS, Management and Control, Control Levels within the protocol architecture have been discussed, and some control strategies to guarantee the needed QoS to the users have been mentioned.

A simple general structure of a distributed control mechanism has been described in detail. The control structure is intended to guarantee QoS and it can manage two types of control: a local and an end-to-end one. The proposed scheme has been and is currently being tested by using multimedia applications in a real environment.

References

[1] A.Miller, "From here to ATM", *IEEE Spectrum*, June 1994, pp. 20-24.

[2] A.Campbell, G.Coulson, D.Hutchinson, "A quality of service architecture", *ACM Comput. Commun. Rev.*, vol. 24, pp. 6-27, April 1994.

[3] F.Davoli, M.Iudica, A.Lombardo, M.Maresca, S.Palazzo, "A network management and control scenario for multimedia applications", *Annales des Télécommunications*, vol. 49, no. 1-2, pp. 65-74, Jan. 1994.

[4] L. Crutcher, A. A. Lazar, "Management and control for giant gigabit networks", *IEEE Network*, vol. 7, no. 6, Nov. 1993.

[5] K.Malinowski, "Practical issues of coordination in control and optimization of large-scale stochastic systems", in S.G.Tzafestas, K.Watanabe, Eds., *Stochastic Large-Scale Engineering Systems*, Marcel Dekker, Inc., New York, 1992.

[6] W. Stallings, *Local and Metropolitan Area Networks*, 4th Ed., MacMillan Pub. Co., New York, 1993.

[7] P. V. Rangan, H. M. Vin, S. Ramanathan, "Communication architectures and algorithms for media mixing in multimedia conferences", *IEEE/ACM Trans. Networking*, vol. 1, no. 1, pp. 20-30, Feb. 1993.

[8] J. Escobar, L. Partridge, D. Deutsch, "Flow synchronization protocol", *IEEE/ACM Trans. Networking*, vol. 2, no. 2, pp. 111-121, April 1994.

[9] C. Douligeris, I.J. Pereira, "A Telecommunications quality study using the analytic hierarchy process", *IEEE J. on Select. Areas in Commun.*, vol. 12, no. 2, pp.241-250, Feb. 1994.