Teletutoring over a Trans-European Broadband Network

Contacts:
Yu-Hong Pusztaszeri (EPFL)
yhp@tcom.epfl.ch
Philippe Dubois (EURECOM)
tel: +33 93 00 26 44
Fax: +33 93 00 26 27
dubois@eurecom.fr

June 8, 1993

Version 3.0

Abstract

We will describe the multimedia teletutoring environment jointly developed by EPFL and Eurecom in the context of the first 34 Mbps Trans-European ATM network interconnecting sites in France and Switzerland. This network was called the Broadband Exchange over Trans-European Links (BETEL). The aim of this report is to describe the BETEL teletutoring platform, scenarios and building blocks, together with performance evaluation, limitations and future enhancements of this prototype. Focus is placed on the interactive audio and video communications part of the application.

1. Introduction

The trend in today's telecommunication networks is migrating towards Broadband Integrated Service Digital Networks (B-ISDN) to support integrated high-speed data, voice and video communications. Asynchronous Transfer Mode (ATM) is the packet switching and multiplexing technique chosen for B-ISDN to provide services with different Quality of Service (QoS) requirements. Meanwhile, new video and audio coding standards are emerging, and many commercial products, both hardware and software, are now available to integrate audio and video with conventional digital data communication.

With this in mind, the European Parliament launched the DIVON program (Demonstration of Interworking Via Optical Networks) in 1992 to prepare and promote ATM technology and new B-ISDN services. The BETEL project, funded by the European Commission and the Swiss Government, was one of the four projects in this program. The aim of BETEL was to run user driven applications over one of the first 34 Mbps international ATM networks.

Two innovative applications were designed to satisfy specific user needs and were demonstrated over the BETEL platform (see Figure 1). The first application, teletutoring, involved interactive multimedia communications between students at the Institut Eurecom in Nice, France and a teacher at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. The other was a meta-computing application for sharing computer resources between the European Laboratory for Particle Physics (CERN) in Geneva and the National Institute of Nuclear Physics and Particle Physics (IN2P3) in Lyon [1, 2].

The goal of this paper is to give an overview of the BETEL teletutoring application. Section 2 describes the BETEL teletutoring platform, scenarios and building blocks, and section 3 outlines the user interfaces while section 4 is devoted to an interactive audio and video communication tool developed for this prototype and section 5 contains brief description of echo cancellers built for this experiment. Section 6 gives detailed performance studies of this prototype. Finally, section 7 discusses the limitations and future enhancements of the system.

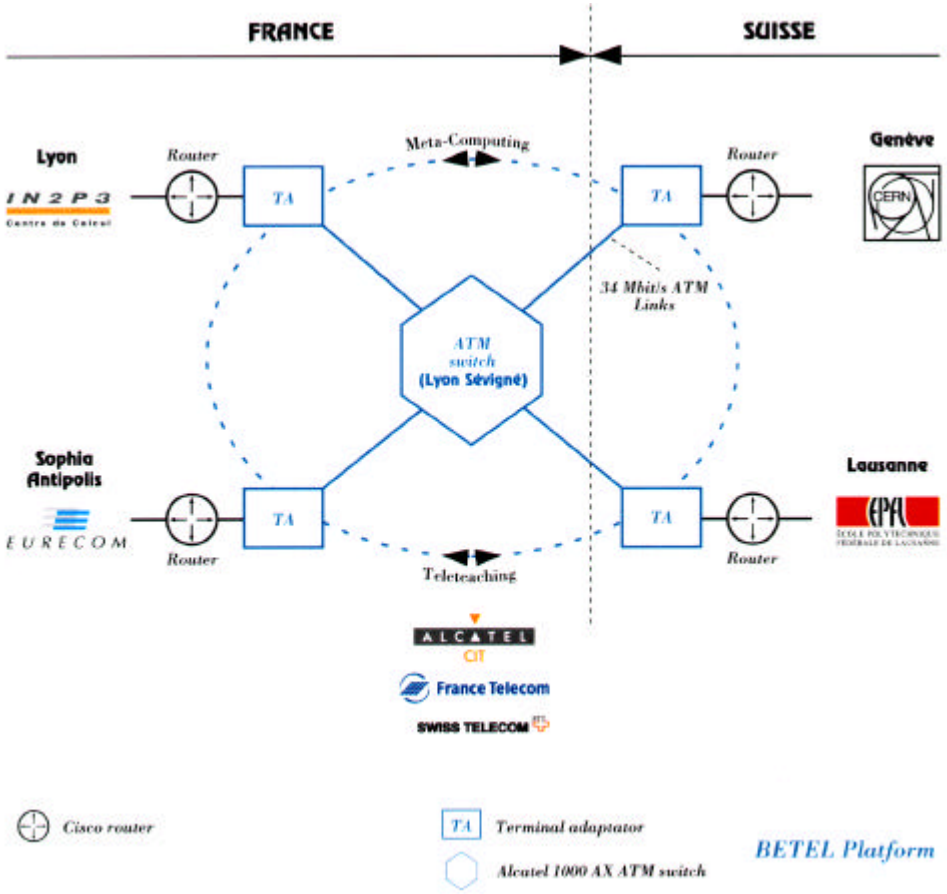

Fig. 1. BETEL: Europe's first operational ATM network

2. An Overview of the BETEL teletutoring application

2.1 BETEL teletutoring network infrastructure

The BETEL network infrastructure shown in Figure 1, is based on ATM technology, and supports FDDI LAN interconnection, ATM transfer service, and AAL 3/4 data service [3]. The FDDI LANs at EPFL and Eurecom were interconnected to the BETEL network by means of Cisco routers.

The endsystem protocol stack was imposed by Cisco routers. Thus the standard Internet protocol suite and FDDI protocols were used. In the BETEL teleteaching application, UDP was responsible for end-to-end real-time audio-video transport service during the videoconferencing session, while TCP was used for data connections between the shared sessions. The user data was encapsulated into IP datagrams, then into FDDI frames, and finally directed to the appropriate remote hosts via FDDI interfaces and Cisco routers, the ATM terminal adapters and the ATM cross-connect.

The Cisco model 7000 is a high performance multiprotocol router. It has Ethernet, Token Ring and FDDI LAN interfaces, and serial interfaces such as the High-Speed Serial Interface (HSSI), used to bridge LANs to high-throughput WANs, i.e., Switched Multimegabit Data Service (SMDS). This router also converts IP addresses into E.164 address numbering scheme.

The ATM terminal adapter provides interfaces between HSSI and AAL3/4 protocols and supports the connection oriented data service. It validates E.164 source addresses, maps the E.164 destination address onto a virtual channel connection (both VP and VC), and provides group addressing such that data can be multicasted to several destinations via separate VPs. The adapter also provides cell header generation and validation, cell rate adaptation and ATM line interface. In addition, it also supports virtual channel and virtual path related functions, such as traffic shaping and VCI / VPI allocation. The BETEL traffic matrix [4] defined by the terminal adapters is given in Table 1.

	CERN	IN2P3	EPFL	EURECOM
CERN		18.5 Mbit/s	2 Mbit/s	2 Mbit/s
IN2P3	18.5 Mbit/s		2 Mbit/s	2 Mbit/s
EPFL	2 Mbit/s	2 Mbit/s		18.5 Mbit/s
EURECOM	2 Mbit/s	2 Mbit/s	18.5 Mbit/s	

Table 1. BETEL traffic matrix

Finally, the ATM cross-connect supports ATM line interface, VP multiplex / demultiplex and VP switching. The end-to-end protocol stacks are shown in Figure 2. In BETEL, only connection oriented data service and point-to-point ATM connections were implemented, and the multiplexing of several VCs over a VP was not available.

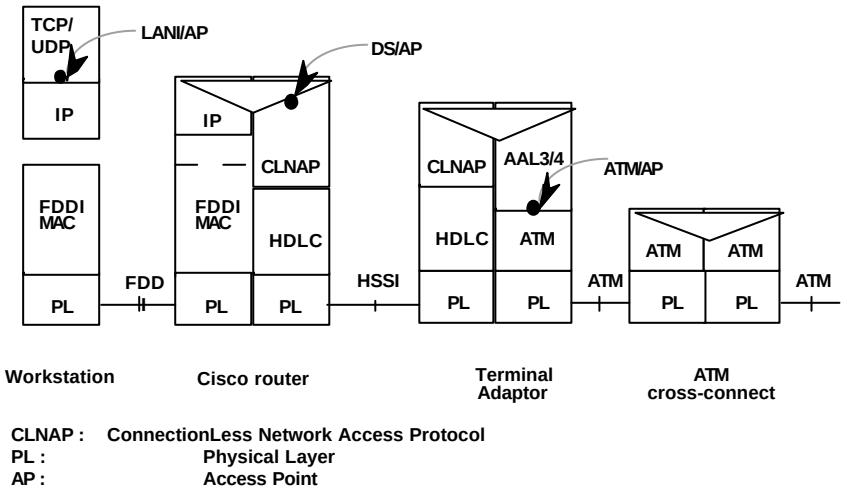

Fig. 2. Protocol architecture

Figure 3 shows the teletutoring network infrastructure. The network topology at Eurecom is more complex than at the EPFL site. Not only an FDDI ring was dedicated for high speed audio and video transmission, but also an Ethernet was used to connect student workstations

to the BETEL platform, and to support connection control and shared workspace data communications at Eurecom. Because the shared workspace sessions generate relatively low data rate, it was not required to connect student workstations to the FDDI ring.

Moreover, the distribution of audio and video signals in the classroom used an analog audio/video switch. All cameras, microphones, monitors, and loudspeakers were connected to the switch, which was controlled by a dedicated software driver to establish and release audio, video and data connections. Using existing analog infrastructure at Eurecom provided a cheap solution since no video compression hardware was needed in each student workstation.

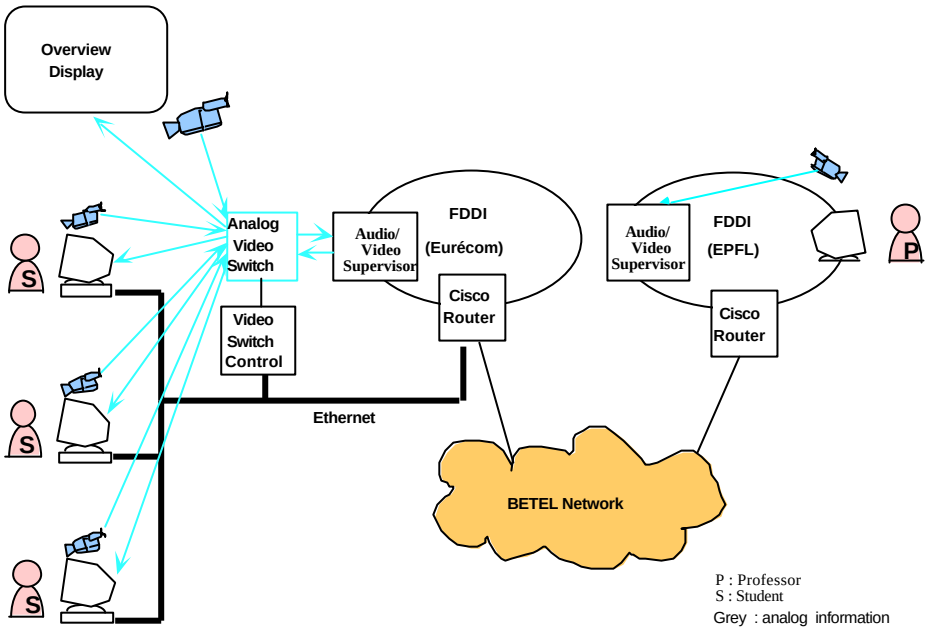

Fig. 3. Teletutoring network infrastructures at EPFL and Eurecom

The teacher at EPFL may either receive images from a camera at Eurecom which gives the global view of the classroom or see a collection of students images through their individual cameras (not through the global camera) using a Picture-In-Picture (PiP) device. The PiP output is connected to the audio-video switch and its inputs come from private cameras of the student. In addition, the image and voice of the teacher may be sent to the global monitor and global speakers in the classroom when he addresses the whole class. On the other hand, the teacher can be engaged in a private conversation with a student while his image is sent to the global monitor and his voice is delivered to the student only. Moreover, a student may send video images from his private camera to his own monitor so that he may see that his own image is adequately transmitted.

2.2 Teletutoring ergonomic

The design of this teletutoring environment was based on a user centered methodology [6], where emphasis is put on the users experiences in relation to the human-computer interaction. This intuitive and user friendly environment enables people at different locations to communicate easily and interact naturally. The user interface and setups of teletutoring classroom and teacher's office were carefully designed and tested (see Appendix A). The

BETEL demonstration room setup at Eurecom and EPFL were just an extended version involving telepresentation facilities.

In both locations, several large TV monitors were used to display images of the audience, of the teacher and students received from the other site. Special attention was also paid to the room acoustics. A set of loudspeakers were placed carefully and echo cancellers were used at both sites to reduce echoes and ensure adequate audio quality. In addition, a large overhead projector was used to display slides during the presentation and show the workspace of the student or teacher while the teletutoring demonstration was taking place. Moreover, each teletutoring unit was equipped with a workstation, a TV monitor, a video camera and a microphone. During the teletutoring session, a student or teacher used one of such units. Figure 4 shows the physical arrangement of the unit in the teacher's office.

Figure 4. Teletutoring setup in the teacher's office

2.3 BETEL teletutoring demonstration

Using this teletutoring infrastructure, the BETEL teletutoring demonstration was successfully carried out at EPFL and Eurecom in December 1993. Following the joint presentation of the project, teletutoring scenarios were demonstrated.

During the presentation, the presenters at Eurecom and EPFL took their turns to present parts of the project. The slides were shared and displayed simultaneously using shared workspace tools. When a presenter at one site was speaking, his images were sent and displayed on the global monitor at the other site while the images of the audience, which was not physically present in the same room as the speaker, were shown on the global monitor so that the speaker could see his audience (both local and remote) and the remote audience could always see the speaker. Similar techniques were also used to allow interactions between audiences from different locations.

In the eight minute run-the-show part of the demonstration, the teacher at EPFL first greeted his three students at Eurecom and the audiences at both sites. Then he checked how his students were progress in their work by establishing audio-video connections to each of them. When he received the "help" signals from his student, he re-established connections to him and asked him to send his workspace so that he could work with him on the problem. This teletutoring scenario is summarized in Figure 5.

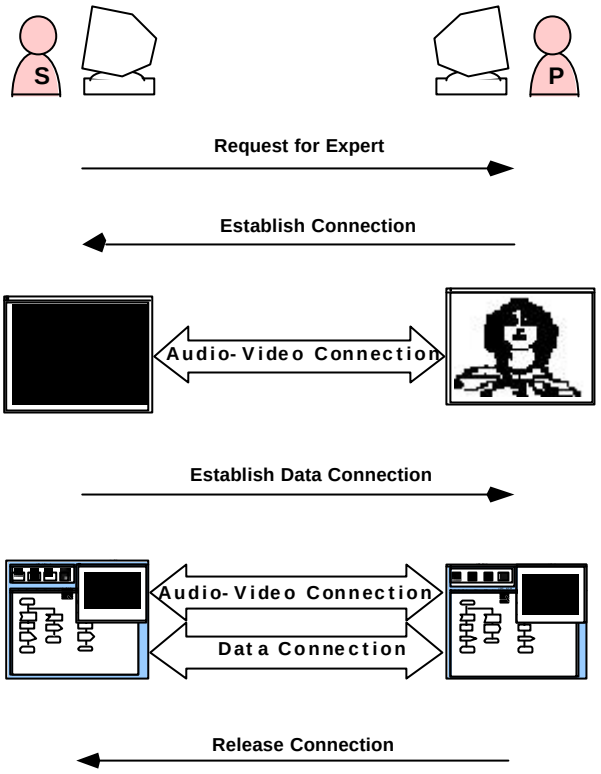

Fig. 5. Teletutoring Scenario

2.4 Building Blocks

Owing to the time constraint of one year (January to December 1993), the choice of the platform was confined by the hardware and software then available on the market. The building blocks are summarized as follows:

- Hardware:
  - HP 9000/700 workstations
  - SUN Sparc10 stations
  - Parallax video acquisition board
  - Echo canceller

- Software Modules:
  - User Interface
  - Connection Control
  - Shared Workspace Manager
  - Audio - Video Supervisor

Commercial workstations were used to build this prototype. Sun Sparc10 stations equipped with Parallax boards were used for audio and video acquisition and transmission, while Hewlett-Packard (HP) workstations were used as workspaces which could be shared between the teacher and the students using SharedX, a Shared Workspace Manager provided by HP.

The User Interface was developed on the HP platform to provide an intuitive and user friendly access to the Audio-Video Supervisor (AVS) and Shared Workspace Manager. On the other hand, AVS was implemented on the Sun platform to provide real-time audio/video acquisition and transmission functionality. Moreover, echo cancellers were used to reduce echoes generated in the BETEL teletutoring network.

3. User interface and connection control

3.1 Functionality

The user interface of the teleteaching application had been designed according to the specifications in [5]. The main functionality of the interface are :

High level audio/video connections control : The connections are related to the state of the interaction, under the control of the teacher. The students are not able to modify the state of the interaction. The state can be either global (the teacher speaks to all students) or local (the teacher speaks to one specific student). Both the student and the teacher always know which state (global or local) they are currently in. The teacher has a list of students whose workstations are connected to his and he can be engaged in a private conversation with one of the students. The teacher has also a global button in his user interface (see Figure 7) for establishing audio and video connections with the class.

Audio-video output device control : The user interfaces has buttons which allow the users to control audio and video outputs. The student, for example, may see himself in his local monitor, and may also mute his audio device. The teacher can view the classroom either from a global camera or as a collection of images, one for each student, using the PiP service, or he can choose a global view of the classroom with a superimposed image of a single student.

Shared workspace connection control : During global communication, the teacher may display his windows on the global monitor acting as a whiteboard or on each student's workspace, while during private communication the teacher and the student can share their windows with each other. In each case, the user has to select the window to be shared by clicking in it. The control software keeps track of the window sharing status. For instance, when changing from local to global state, all the shared windows are unshared, but when the teacher re-establishes the private connection with the student all the previously unshared windows can again be shared.

Student question support : The students may send questions to the teacher, for example, to ask for a private connection with the teacher. The user interface for students and the teacher are shown in Figures 6 and 7. The teacher can access a chronological list of all received questions. The teacher answers one question at a time. The students receive feedback on the relative position of their question on the list.

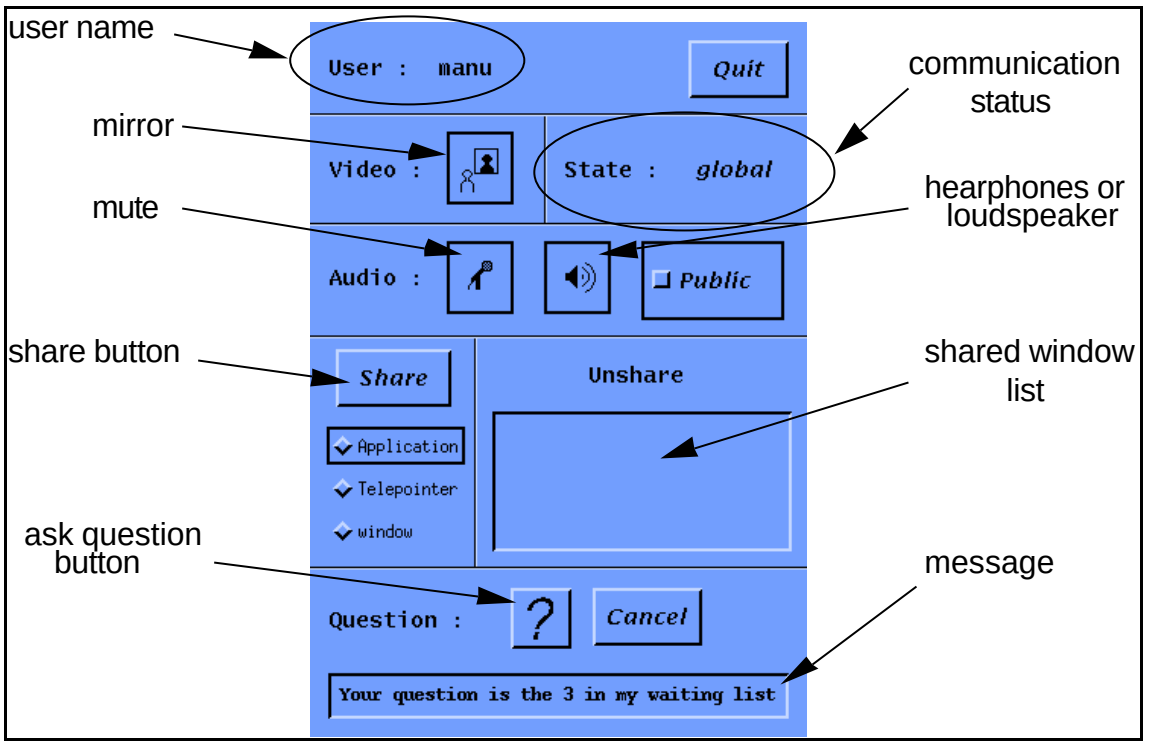

Fig. 6. User interface for students

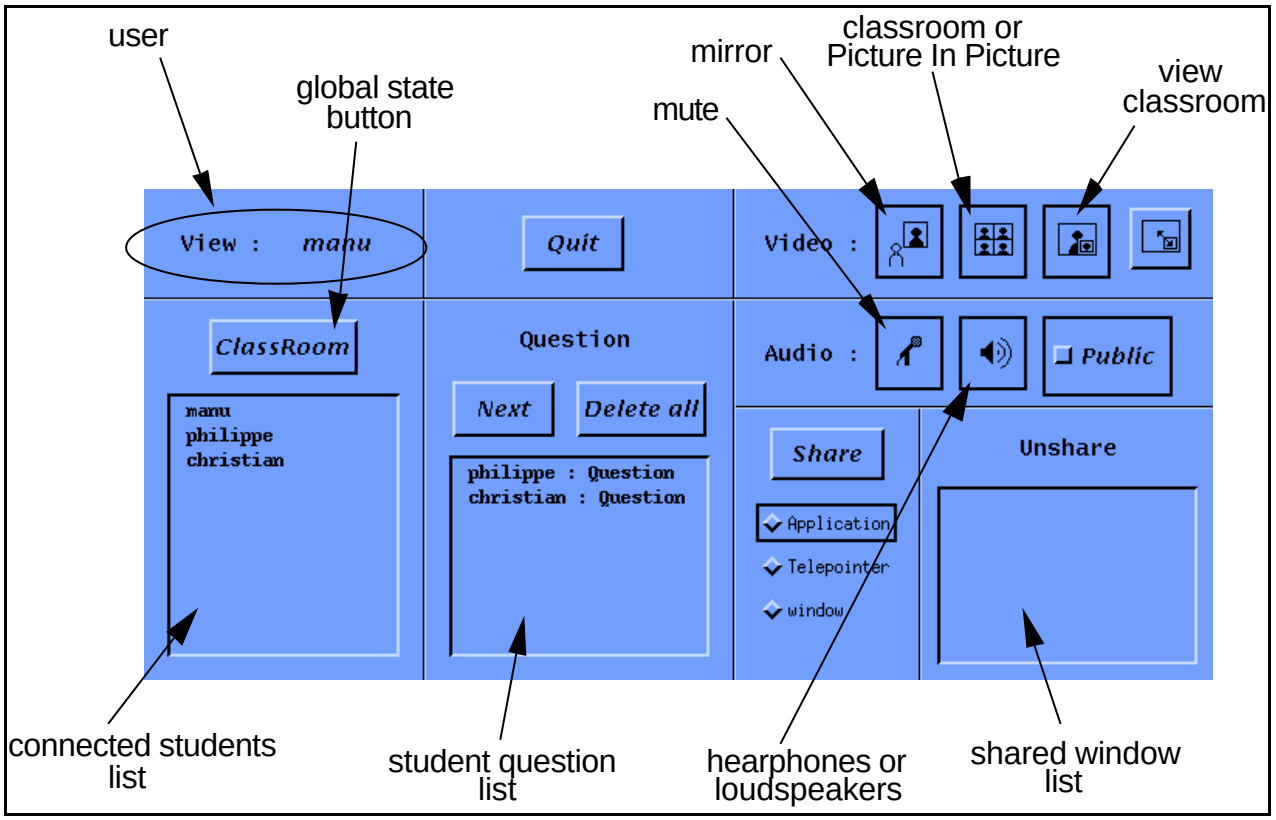

Fig. 7. User interface for the teacher

3.2 User interface software architecture

The user interface software has a client-server architecture as shown in Figure 8. The server is called the Session Agent, and the clients, the user interface, are called Teacher interface (Figure 7) and Student Interface (Figure 6). The Session Agent is an intelligent process which knows the topology of the audio-video analog network and digital data network (i.e. the names and locations of each device) and it is also responsible for managing all audio, video, and shared workspace connections. It receives high level commands from the user interfaces, for instance, "*set communication state of teacher and student Tom to local* or *set communication state of all to global.*" According to the physical topology of the audio-video analog network and digital computer network, the Session Agent transforms these commands into a set of low level actions, such as "*connect camera1 with monitor2* ...,etc." The actions are effectively handled by two "low level" drivers : an audio/video connection handler, which drives the audio-video switch, and a shared window handler based on X Windows.

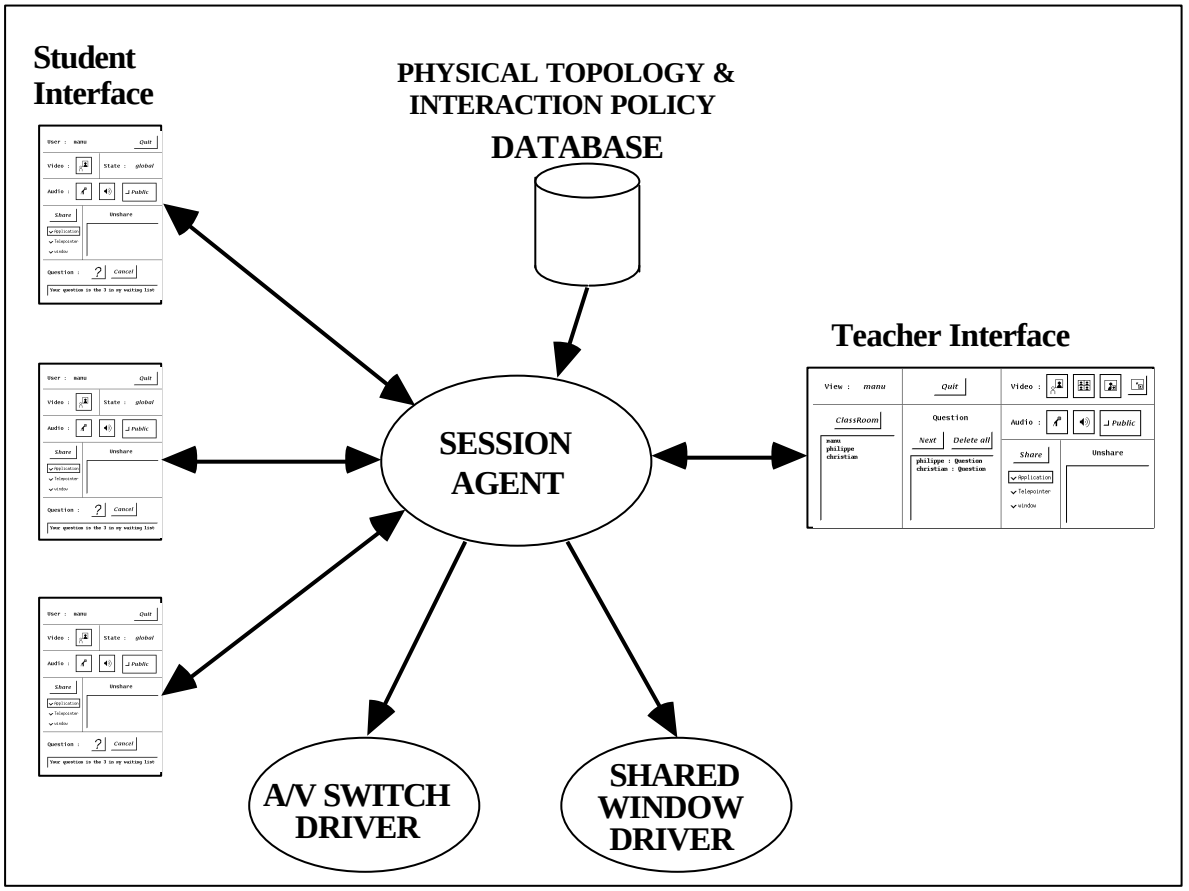

Fig. 8. User interface software architecture

When a teleteaching session is started, the Session Agent should also know the interaction policy of that session. The policy is based on commands stored in a configuration file, for example, "*If mute then disconnect my microphone*." Moreover, the Session Agent is able to manage several parallel sessions, and each session may have its own interaction policy. The Session Agent can manage concurrently several teleteaching sessions, for instance, point-to-point and multipoint videoconferences.

4. Audio - Video Supervisor

AVS provides real-time audio and video acquisition and end-to-end transmission. Under the supervision of AVS, audio and video signals from analog sources are digitized and encoded, and then are transmitted to a remote station via the BETEL network. At the receiver end, the data are decoded, and video images are reconstructed and displayed while audio is being replayed. Audio and video signals are handled in a similar fashion (see Figure 9).

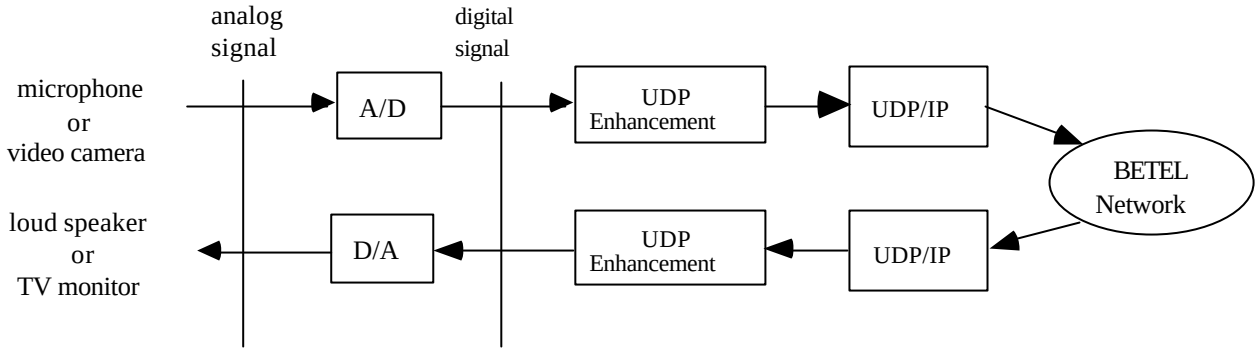

Fig. 9. AVS processing pipeline

4.1 Video acquisition

The current AVS implementation used the Parallax XVideo board. This board was the only hardware available on the market that permitted real-time video compression and decompression at a reasonable frame rate. Details about this board and its performance are given in [7].

The Parallax board can handle analog video input and output in various standards (PAL and NTSC) and formats (Composite, super VHS, YUV and RGB). The video signals are first digitized, then compressed by the XVideo board based on the JPEG standard before being sent to the network. On the receiver side, the digital video signals are decompressed, converted to the analog signals, and displayed on the receiver's monitor.

The programming interface that came with this board uses an extension of the X11 library called XVideoToolkit. Its functionality is exploited by AVS through an extended X-Window server, which provides access to Parallax's graphical accelerator and frame buffer, and supervises video digitization and compression / decompression.

One of the major drawbacks of this board resides in digitized video images, which have to be first stored in frame buffer of the XVideo then compressed by the JPEG Image Compressor of XVideo board. Hence, video images cannot be compressed without being first displayed locally.

4.2 Audio acquisition

Audio streams are digitized, recorded and played by the SpeakerBox of the Sun Sparc10 station. The SpeakerBox audio peripheral provides an integral monaural speaker and microphone, stereo line in / out and headphone connections. This SpeakerBox supports different audio qualities and encoding techniques. Moreover, it has a programmable audio device interface.

4.3 Networking issues for audio/video transmission

Real-time audio and video transport service imposes several performance requirements on the network. Since both audio and video sources produce continuous data streams, not only do their temporal relationships have to be satisfied, but they also require large network bandwidth. In summary, interactive audio and video data generated by the teletutoring application impose the following network requirements:

- guaranteed high throughputs
- bounded end-to-end delay and delay jitter
- low loss and error rates
- connection-oriented service, i.e., in sequence delivery
- support for real-time data service
  - higher priority for real-time data
  - selective discard data according to their priority in case of congestion
- synchronization
  - intra-medium
  - inter-media
- adaptive rate-based flow control

The transport layer protocol is restricted to TCP and UDP since IP was imposed by the Cisco router. TCP provides reliable end-to-end connection oriented transport service while UDP and IP support best effort services based on connectionless techniques. The Internet protocol suite was designed for point-to-point non-real-time data service and has many difficulties to meet the network performance requirements demanded by the teletutoring application.

TCP/IP is unsuited for networks with large bandwidth-latency products. The BETEL network is one such networks. The sliding-window flow control with credit allocation does not allow to use the full bandwidth of the BETEL platform. Retransmission in TCP significantly increases end-to-end delays and is unsuitable for interactive audio and video data transport service. Hence, in this context, window based flow control and error control mechanisms found in TCP create problems for real-time audio and video transmission.

On the other hand, the lack of retransmission in UDP makes it a better candidate to transport real-time audio and video data. Since UDP does not guarantee in sequence delivery and does not have any error or flow control, some endsystem enhancements added to UDP are needed. For instance, video frames in general are larger than UDP datagram limit (9 Kilobytes), thus video frames need to be segmented into smaller frames before being sent to a UDP socket and be reassembled together at the receiver end. In order to make the reassembly process efficient, missing frames and out of ordered frames have to be detected.

The minimum UDP enhancements are loss detection and packetization (including segmentation and reassembly for video frames). Therefore, UDP/IP protocols with endsystem enhancements were used to transport audio and video data.

4.4 Implementation Issues

AVS was designed to guarantee best performance. This was done with as little data movements and copying as possible. Only minimum UDP enhancements were implemented. Data were packetized (segmented if necessary) and sent to the UDP socket without any buffering and copying. A small header was added to each packet. The frame sequence information (for loss sequence detection and sequence check) was put in the header, and so were the segment number and total segments in a given video frame also included for video frames. A typical video packet size was about 4 Kilobytes and the minimum MTU[1] in the BETEL network (excluding the Ethernet segment) was also 4 Kilobytes, whereas audio frames had 128 audio samples to ensure low delay and loss rate.

The audio quality was the most important factor in the design of this teletutoring prototype. Voice is still the most common and effective means of communication, although eye contact and other facial information are also important. Hence, some steps were taken to ensure good audio quality. The use of smaller audio frames was one. Another was the use of high quality sound equipment. The Sun microphone did not give as good audio quality as did the Macintosh microphone, so the latter was used, together with semi-professional microphones (connected to line in port of the SpeakerBox). In addition, echo cancellers were necessary to reduce echoes generated by the large round trip delay in the BETEL network. Further, it was impossible to use CD quality audio supported by the SpeakerBox because the echo canceller used could not treat audio with sampling frequency higher than 8KHz. Therefore, audio was restricted to telephone quality. Audio data in AVS were not compressed to reduce processing delay and to ensure good audio quality, as the BETEL network bandwidth was no bottleneck in this application.

Another challenge consisted in justifying the 34 Mbps high-speed links for this teletutoring experiment. Why do teletutoring applications and other interactive multimedia application need broadband networks nowadays? The most obvious answer is video quality. High quality digital video signals demand large bandwidth. Since the XVideo board is one of the most performant real-time compression hardware available, understanding the parameters influencing the video quality and the performance of the XVideo board is essential. Performance studies in section 6 show that the performance bottleneck is in the endsystems.

In order to overcome this bottleneck, a Sun Sparc10 station was dedicated for either to transmit or to receive video sequences. As AVS consists of four independent processes, each process is dedicated to either receiving or sending video and audio data. Transmission of audio and video is hence independent. Therefore, four Sparc10 stations equipped with Parallax boards were involved in the video acquisition and transmission of this prototype.

[1]Maximum Transfer Unit

5. Echo cancellation through adaptive filtering

5.1 Acoustical echo and Larsen effect

When a person at location A speaks, the sound is transmitted to location B by the speaker. With the unavoidable reflections on the walls in that room, shown in Figure 10, a part of the signal diffused by the loudspeaker in B will be taken by the microphone at B. The person at A will therefore hear his own voice generated by acoustical echo. The echo effect becomes disturbing only if the progagation time exceeds a few tens of milliseconds.

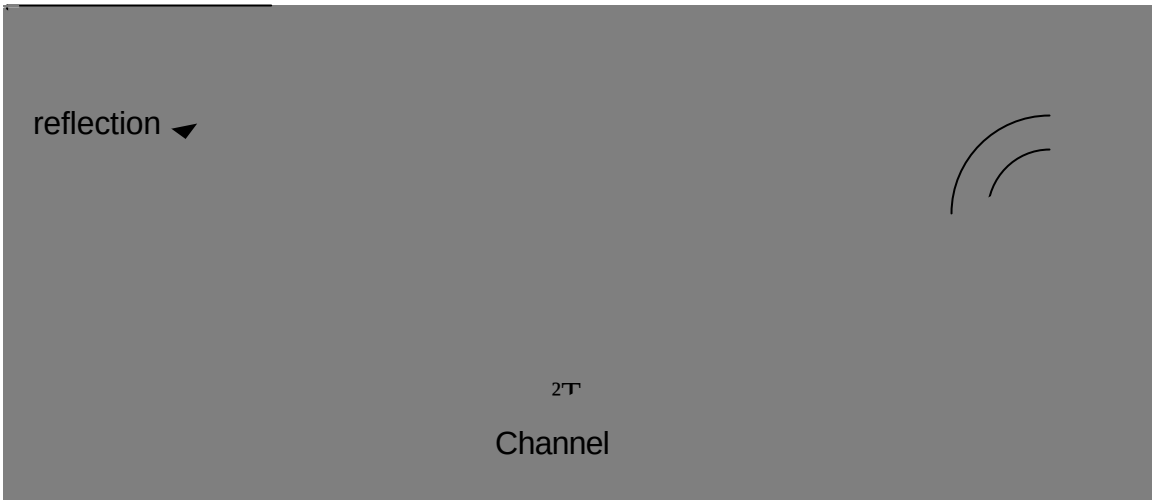

Fig. 10. Acoustic echo path A-B-A

In fact, the signal loops infinitely in the path A-B-A. It becomes a feedback system which oscillates at one or more frequencies. This is called the Larsen effect. This effect is a consequence of an echo. Thus, canceling echo also suppresses this effect.

5.2 Parameters influencing echo

A number of parameters are important to reduce echoes. First, the choice and location of the directional transducers (microphone and loudspeaker) are important. For example, placing directional loudspeakers in an area of low sensitivity to the microphones may reduce echoes. Secondly to that is room acoustics. Reverberation decreases as the room volume increases and the wall absorbtion factor increases.

However, these two approaches are not sufficient to eliminate the echo effect. Thus, an echo cancellation system is needed, implying that at least one echo canceller must be placed at both the transmitting and receiving end.

5.3 Echo canceling through adaptive filtering

The adaptive filtering technique [8] used the Least Mean Square (LMS) algorithm, which has the advantage that no prior knowledge of the room impulse response is required. The adaptive filtering shown in Figure 11, consits of two distinct steps:

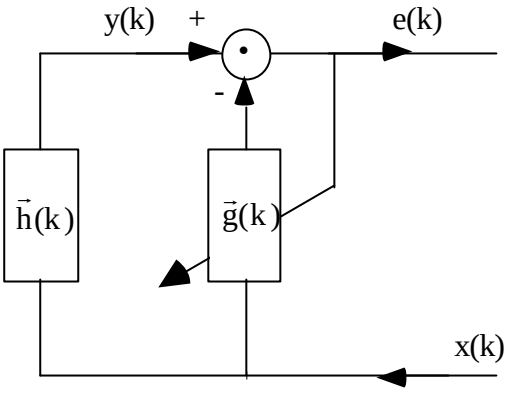

Fig. 11. Feedback of e(k) on the filter coefficients with emphasis on the adaptation process

- estimating an filtering error

$$e(k+1) = y(k+1) - \vec{g}^{T}(k) \cdot \vec{x}(k+1) \qquad (1)$$

- updating the coefficients $\vec{g}(k)$, using the error e(k+1).

$$\vec{g}(k+1) = \vec{g}(k) + Ke(k+1)\vec{x}(k+1) \qquad (2)$$

Where

$$\vec{x}(k) = \begin{bmatrix} x(k) \\ x(k-1) \\ \cdots \\ x(k-N+1) \end{bmatrix} \quad \vec{g}(k) = \begin{bmatrix} g_0(k) \\ g_1(k) \\ \cdots \\ g_{N-1}(k) \end{bmatrix}$$

x(k) represents the sample at instant k,
$\vec{x}(k)$ represents the vector of the N most recent samples at time k
$\vec{g}(k)$ is the vector of the N filter coefficients at time k
y(k) is the sample coming from the microphone at time k
e(k) is the error at time k
K represents the adaptation step

The convergence time and stability of this system depend heavily on the value of the adaptation step K, which depends on the length of the filter N and on the input signal power. For more details on the selection of K and this algorithm please read Appendix B.

6. Performance evaluation and multimedia traffic analysis

6.1 Measured TCP/IP and UDP/IP performance

The purpose of this study is to estimate the upper bound in performance which is available for applications running on top of TCP and UDP. The TCP/IP and UDP/IP throughputs were measured in both local FDDI environment and on the BETEL teletutoring platform. Details of this study are in Appendix C.

Figure 12 shows the measured performance in the local FDDI LAN environment, where UDP/IP (without UDP checksum) could achieve 7.3 Mbytes/s throughput for a message size of 4K bytes but with a maximum of 20% losses while TCP/IP has a maximum throughput of 5 Mbytes/s for the same message size.

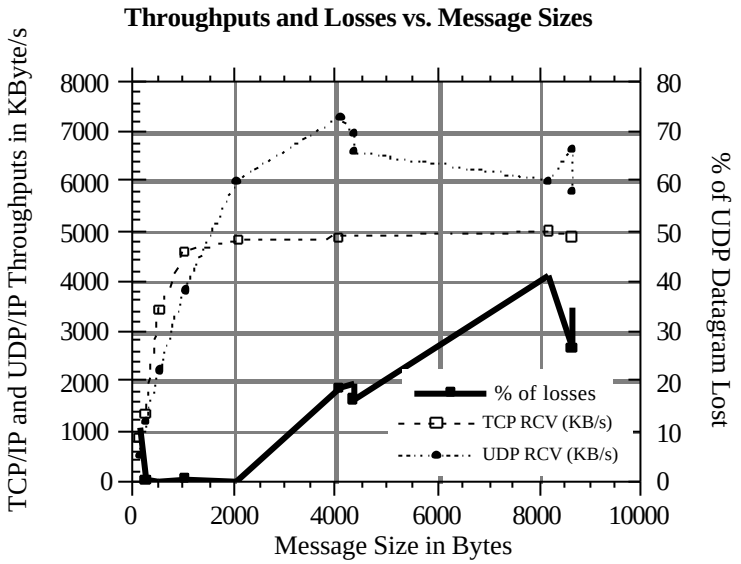

Fig. 12. Performance measurements between two Sparc 10 stations over FDDI

On the other hand, TCP/IP throughputs stabilize at around 1.05 Mbyte/s in BETEL teletutoring network for message sizes above 2000 bytes, as shown in Figure 13. Moreover, the Round Trip Time of the EPFL-Eurecom BETEL links is about 12 milliseconds for 64 byte messages and 17 milliseconds for messages of 1024 bytes.

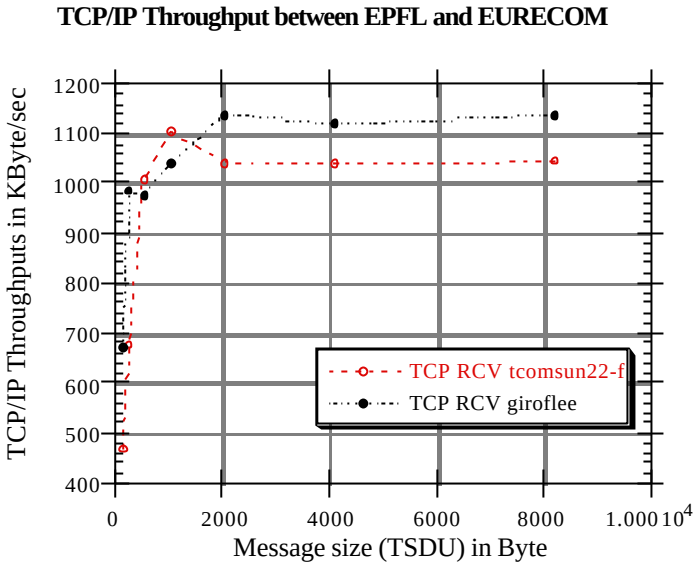

Fig. 13. Maximum TCP/IP throughputs in the BETEL teleteaching platform

These results show that UDP can attain higher throughputs than TCP, but UDP suffers from losses. In addition, the BETEL teletutoring network has a large bandwidth-latency product. Hence, TCP is not suited for real-time audio and video data communications and the alternative is thus to use UDP. However, since UDP is unreliable, some UDP enhancements are necessary. Moreover, since UDP can obtain higher throughput than TCP, there is at least a 1.05 Mbyte/s bandwidth for real-time audio and video data transported by UDP across the BETEL teletutoring network between EPFL and Eurecom.

6.2 Measured video performance

Video communications demand large bandwidth and its peak performance is likely to be limited either by the network or by the endsystem itself. Understanding the parameters influencing video performances can help us to obtain the best video quality. By measuring video bit rates and frame rates, we can gain an objective insight into video performance and hence identify the bottleneck of this prototype. The performance measurements in this study were taken between two Sun Sparc10 stations in an FDDI LAN at EPFL.

First, video bit rate depends on the Q factor which is used by the XVideo board to determine the quantization level and to control the compression factor. The higher the Q factor and the larger the compression factor, the lower the video bit rate and video quality. Another important factor is the video frame rate which is the number of video frames captured and played per second. Figure 14 shows the relationship between the unidirectional video bit rates and the number of frames captured per second when only a quarter of PAL resolution was used and the Q factor was at 50. With the rate control mechanism, the video bit rate increases proportional to video frame rate.

However, when the rate control mechanism is disabled, we can push the system to its maximum capability, reaching 34 video frames captured per second at a video bit rate of 5.7 Mbit/s. Similarly, transmitting the full PAL resolution of video images with the same Q factor,

the system can reach only 12 frame per second. Moreover, the measured results from section 6.1 shows that the bandwidth available for video communications in BETEL teletutoring platform is larger than the maximum video bit rate which the endsystem can deliver. Therefore, it is clear that the bottleneck lies in this system and not in the network.

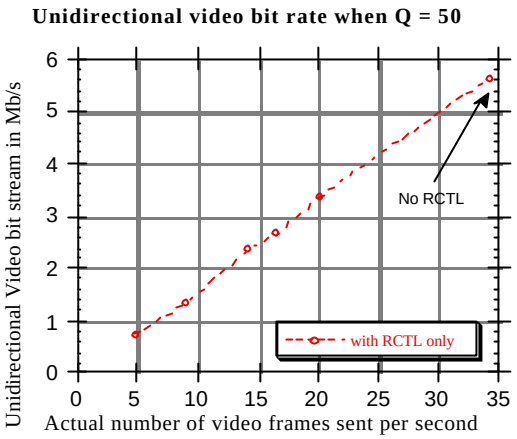

Fig. 14. Video bit rate vs. video frame rate

6.3 A theoretical performance model of the multimedia workstation

In order to understand the performance limitation of this prototype, a theoretical performance model of the multimedia workstation is used to precisely identify the bottleneck in the endsystem, that is, a Sun Sparc 10 model 51 (50 MHz, 64 MB RAM, 1 GB Disk) equipped with the Parallax XVideo acquisition board and Sunlink FDDI/S interface. A basic assumption is that only two-party video conferences will have to be established.

The architecture of this multimedia workstation is outlined in Figure 15, which illustrates that for each video connection, data had to cross the system bus twice (from the Parallax frame buffer to the system memory and from the system memory to the FDDI interface), that is, four times for a two-party video conference. The bottleneck seems to be the SBus, as the MBus is twice faster [9].

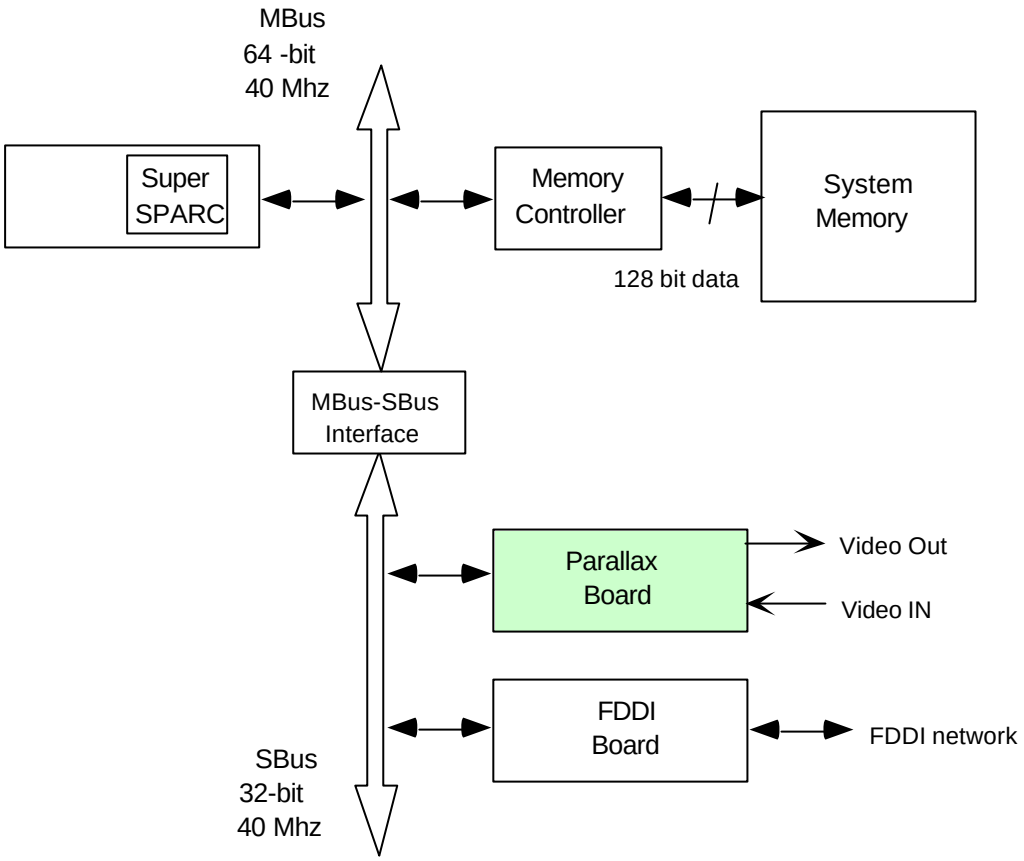

Fig. 15. SPARCstation 10 architecture

In order to quantify the workstation performance, a simple theoretical model presented in [10] was used. In this model the service time, namely the time taken by an individual server to process a packet, is broken into two parts, fixed service time and incremental service time. A fixed per-packet service time includes the time to take to filter packets by the network interface, and the time spent in datalink layer, network layer and transport layer processing, interrupt processing, memory management and context switching. On the other hand, an incremental service time that varies with packet size consists of the time needed for data movements between host main memory and network interface, between host main memory and video board, between system memory and user memory, and optionally the error checking overhead. Finally, the throughput (number of packets per second) is defined as follows:

$$\text{Throughput} = \frac{1}{\text{fixed service time} + \text{incremental service time}}$$

6.4 Compare the theoretical and measured results

Applying this model, the theoretical throughputs of raw data and unidirectional video data transfers using UDP/IP can be obtained (Appendix D). These theoretical approximations are compared with the measured results in sections 6.1 and 6.2.

The UDP raw data throughput expressed in function of the packet size is an expression of the equation (4) where p is in packet size in bytes. Thus the slope at its origin is the inverse of the

fixed service time. Figure 16 compares the theoretical and measured maximum raw data transfers between two Sun Sparc 10 stations interconnected by an FDDI ring. The good match of the two curves for small packet sizes show that the estimated fixed service time of 200 µs is very accurate.

$$f(p) = \frac{p}{200 + 45\,p} \qquad (4)$$

For large packets, however, the difference between theoretical and measured results becomes significant. It is possible that the effect of DMA transfers on the CPU (contention for memory) has been under-estimated. The dotted curve on the previous figure (which takes into account the DMA transfer time or, in other words, does not consider it as happening in parallel with normal CPU processing) tend to confirm this explanation.

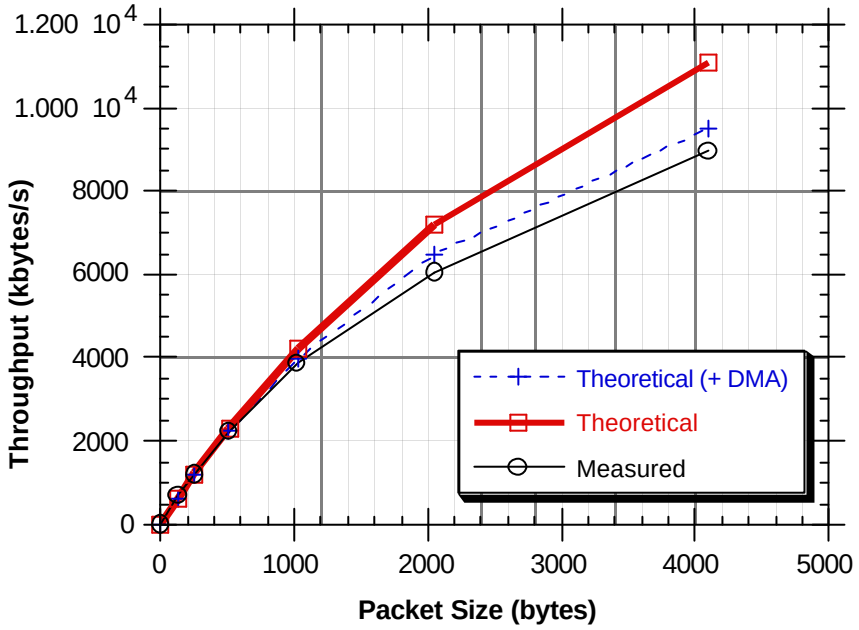

Fig. 16. Comparison of data transfer performances

On the other hand, the theoretical approximation of the unidirectional video data transfer between two workstations. The theoretical performance evaluation may be performed, based on the following parameters:

- image size = 384 x 288 the quarter of a PAL image
- average compressed image size : 20 Kbytes (i.e. five 4 Kbytes UDP packets are needed to transfer one compressed image)
- assumed DMA speed of 50 Mbytes/s

To take into account the overhead introduced by AVS and the video acquisition board during image processing (in particular context switching, X overhead), the fixed service time per packet has been doubled. To allow computation of end-to-end throughput, the incremental

service time which refers to images is expressed in terms of average processing time per packet. The incremental service time required by each sequential operation can be estimated. This gives the theoretical throughput of 6.75 Mbps.

This result is very close to the measured value of 5.7 Mbps. The difference should arise mainly from the variable size of a compressed image which is in general not an integer multiple of the optimal packet size and thus reduces the transmission efficiency.

7. Limitation and future enhancements

The implementation philosophy of BETEL was to integrate currently available technology and build a demonstrator within one year. The teletutoring prototype unearthed several shortcomings in the original design, and inherited the limitations of the current technology. This teletutoring experiment used a hardware dependent, point-to-point configuration (i.e., EPFL-Eurecom), and used the UNIX operating system and the Internet protocol stacks. There was no build-in synchronization and rate-control mechanisms implemented in the videoconferencing system. Therefore, enhancements are needed in the following areas : multipoint and multiplatform teletutoring configurations, and system support for interactive real-time teletutoring applications.

7.1 Hardware dependency

The hardware dependency could be relaxed as the video compression and decompression hardware and shared workspace tools were progressively made available. The release of a Parallax board for the HP platform has been announced for the Spring of 1994. AVS may then be easily ported to the HP platform since the Parallax boards (both for Sun and HP platforms) are using the same C-cube chips which are based on the JPEG compression/decompression standard. In addition, AVS can be also modified to use other video compression hardware, for instance, those based on the MPEG standard when they become available. Moreover, Sun has recently released a commercial product called ShowMe2, a competitor to SharedX. Unlike its predecessor (ShowMe), ShowMe2 can be used to share applications, allowing both videoconferencing and shared workspace tools to be integrated on the same platform. The BETEL teletutoring prototype could then be ported to multiple platforms.

7.2 Audio Quality

Audio quality is most important in teletutoring, and was hampered by echo, which was a serious problem because of large latency audio experienced in the BETEL links. Several echo cancellers were designed to cancel this effect, but these devices could only process one speaker at a time and created problems when used with audio mixing devices. An audio enhancement is to design new echo cancellation algorithms which support CD quality audio and can be used with mixing devices. This would benefit mostly teletutoring with geographically dispersed students.

7.3 Scalability

The current prototype is limited to a one-to-one interaction. The teacher can interact with one student at a time, although his image and voice can be broadcast to everyone in the classroom. Two or more students cannot engage in a discussion. All video and audio signals have to be transported via point-to-point audio and video connections. It would be useful, if either a teacher could simultaneously supervise several students from different sites, or several teachers from different sites could interact together. Thus, a fully meshed digital multipoint videoconferencing is needed.

7.4 System support for teletutoring

One of the long term solutions to the performance problem is to implement endsystem support and network support for interactive multimedia applications, such as teletutoring. A better multimedia workstation architecture is needed to sustain transmission of large amounts of data through the system buses and to minimize data movement and data copying.

The UNIX operating system is not adequate to support real-time services. The clock resolution and scheduler of the Sun OS 4.1.3 illustrate this point. Since its clock resolution is not higher than 20 milliseconds [11], it is difficult to implement any efficient audio-video synchronization mechanisms. In addition, the scheduler cannot give higher priority to real-time media.

The Internet protocol cannot guarantee high throughputs and bounded delay and jitter which are required for teletutoring applications. Thus, network protocols supporting QoS are needed here. Moreover, the current network protocols do not support multicast service which is essential in the multipoint teletutoring configurations. Furthermore, endsystems should at least support audio-video synchronization and adaptive rate-based flow control.

8. Conclusion

With the aide of high quality videoconferencing and shared workspace tools, the BETEL multimedia teletutoring prototype was successfully demonstrated at the end of 1993 over the first 34 Mbps Trans-European ATM network. Echo cancellers were essential in ensuring high quality audio in this experiment. The performance bottleneck of this prototype was at the endsystem level, particularly in the video acquisition board. The UNIX operating system and BETEL protocol stacks provided best effort services which were not ideal but satisfactory to support the BETEL teletutoring application. UDP/IP, not using any explicit audio-video synchronization and adaptive rate-based flow control mechanisms, were used to transport real-time audio and video data. More robust and realistic teletutoring scenarios will be realized in a multipoint and multiplatform environment in the framework of European ATM pilot experiment starting in July 1994, for example, a Europe-wide M. Sc. program with distributed lectures, classrooms and campuses.

Appendix A: Teleteaching scenarios: The Process towards specification.

Teleteaching ergonomic design and user interface specification were based on the user centered methodology. The term “User Centered System Design” was first coined by Norman and Draper [5] in 1986. In user centered system design the emphasis is on users. The ultimate central question that underlies all user centered design is "what does the experience like for the user?" User centered system design methodology does not represent any specific scientific method. It regards the art of system or application design as a combination of a number of disciplines: engineering, history, science and arts. We adopted this approach in the design of the BETEL teleteaching application for the classroom setup and the user interface specification. It ensured us that real life situation and problems would be studied in their full complexity in relation to the human-computer interaction. We followed three principles of design:

Initial design:

Our main goal in designing the teleteaching application was to come up with a system which would be intuitive and easy to operate. Already in the initial stage of the design we identified the target users, professors and students, and observed them in a real situation, TP sessions. Based on our observation, for example, number of monitors and the kind of interactions made, we defined the initial technical requirements of the application and the specific tasks to be performed by the users. The goal at this initial stage was to construct the first prototype which will be flexible enough to allow iterative design.

Empirical measurements:

We selected four professors and eight students to participate in our study. They differed in their learning or teaching style, their experience in using a teleteaching application and their exposure to a multimedia environment. For each prototype testing, we used one professor and two students to perform some tasks. Each participant took part in at least two prototypes testing. They carried out simple tasks and their performance, thoughts and attitudes were recorded and analyzed after each trial. In doing so we took into account the user's progress.

For each trial we conducted, studies involving observation of the behavior and attitude of the users while performing tasks using the different prototypes, i.e. the pattern of interaction between the professor and the students. To understand how they went about their work and what their problems were, we collected the users comments while they were working with the system. This technique is called "thinking aloud" and it is borrowed from the field of cognitive psychology.

At the end of each trial, we conducted interviews and discussion sessions. Users were free to suggest what they did not like and what they thought should be added. The main idea was not to have the users agreeing on the design but rather to create a potential situations whereby the users instilled their knowledge and concerns into the design process from the very beginning. We conducted what is called a "Participatory design": our users became part of the design team from the very outset.

Iterative Design:

When problems were found in our user testing they were fixed tested and redesigned. The iteration was possible only because our implementation strategy permitted early testing of the design feature and easy modification of the evolving implementation.

By using an iterative methods we confronted with the reality of an unpredictable users needs and behaviors that lead us to conduct immediate changes in the design of the application through out the design cycle (for more information please refer to Appendix A).

DEFINING THE FIRST PROTOTYPE

Based on our observation of an actual TP sessions we created and participated in several simulated teleteaching scenarios. Our objective was to develop the first prototype for the physical setup of the professor's office and classroom. Throughout this process, our experience as a professor or as a student was recorded and analyzed at the end of each session.

The following was set to be determined for each scenario :

Interaction Pattern
Professor- Student
Student-Professor
Professor- Students
Students- Professor

Visual/image
Camera(s); number, positions and angles
Monitor(s); number and position
Screen(s); type and size

Audio/Sound
Microphone(s); types number and position
Headphones; type and position
Loudspeakers type, number and position

The data collected determined the following:

Interaction Pattern
The interaction occurs between the **Professor** and the **Students** in two setups: **the classroom** and **the office.**

Two types of audio visual communication links were identified: **Global** for general communication link and **Local** for private communication link. **Global Link** refers to the interaction that occurs between the **professor and the whole classroom Local Link** refers to the interaction that occurs between **a student and the Professor and vise versa.**

The communication dialogue between all parties is divided into four states

Global Background State- A link exists between the classroom and the office but no interaction takes place only an awareness.
Global Teaching State- The professor interacts with the whole classroom.
Global Interactive State- The professor interacts with an individual student within the classroom context.
Local Interactive State- The professor interacts with an individual student on an individual basis

We decided to use the following terminology to refer to:
Global- camera and screen
Local - Camera and screen
Global- Audio connection
Local- audio connection
Sharing software - Local and Global computer connection

The following is our preliminary design Spec. for the first prototype

Classroom Set Up

The ***global camera*** should ***be placed in front*** of the classroom so the professor will have a general view of the classroom which include gesture and face expressions.

Camera and monitor, global and local, must be placed together at the ***same position f***or the purpose of interaction otherwise confusion may occur.

The **Prof. image** should be **constantly displayed** on the global screen.

At his stage we found out that the position of the local monitor and camera need to be further investigated because of the following:

A. When the local monitor and camera were placed in an upright position the student respond positively but it was a problem for the professor. The professor asked to see the actual workspace of the student as if he was standing beside him. This applied also when the share workspace software was used.

B. When placed in a shoulder position (beside the students), students felt that it discouraged interaction between them and the Prof.. The professor on the other hand liked the local view from this position.

Camera Angle

The ***Global Camera*** **s**hould provide a***wide angle global view.*** of the classroom

Number of Monitors

A ***separate local monitor*** should be provided for the local interaction. The computer screen should be used for task excision only. A separation between the task and the interaction is essential.

Audio

The sound has to be localize otherwise spatial confusion may occur. For the **local link t**he sound should be distributed only from the **local speakers.** For the global link the sound should distributed only from the global speakers.

PROFESSOR OFFICE

Number of Monitors
A ***separate local monitor*** should be provided for the local interaction only.

The professor must be provided with two simultaneous views; the whole classroom and the individual student with which he is engaged in a local link. The use of **PIP** is suggested for this function.

Areas for investigation

Earphones:

The use of earphone for local interaction should be further investigated

Audio level

Audio disturbance level should be determined for local and global interaction.

Camera angle and position

The field of view of the Professor's camera need to be tested from the following angles;

Close Up- Students see only the professor face.

Medium Shoot- The student (s) see the professor from the waste up. Only a partial view of the working environment

Long Shoot - The student(s) see the professor within his working setup and room architecture

The shoulder position for a local view is suggested to be further investigated in order to simulate a situation where the Prof. stands right beside the student.

The data collected at this stage served as a frame of reference for the design of the first prototype. Using our prototypes the interactions between the participants and their individual action was analyzed according to the following HCI criteria:

Interaction:
- Awareness
- Gesture
- Conversation

Task Execution:
- Individual Action
- Group Action

Errors:

Human errors
Technological errors
UI errors

The different teleteaching prototypes consisted of the following:

A/V connections
Room Architecture
Camera(s) positions and angle
Monitor(s) positions and size
Shared workspace functionality

FIRST SCENARIO

Based on our preliminary spec. we designed the first prototype to be tested using different scenarios. We will describe one scenario in the following section.

Description

Three subjects participated in the first trial; a professor who was an expert with the concept of the remote learning technology, and two students, one who was novice to the technology and the other who was familiar with it.
The professor was situated in an office equipped with the following; a monitor size 28cm a Mac power book, a Sony camera and a Shared Workspace Software, Timbuktu.

The two students were placed in a classroom, one behind the other. The classroom was equipped with the following; A large screen 28cm for Global link and a small screen 14 cm for the local . The global screen was facing the students in an angle. The local monitor was placed in front of each student beside their computer screen. A global camera was mounted on top of the global screen and the local camera was on top of the local one. Mac terminals served as the work station.

The local audio distribution was via an omnidirectional Microphone for one student and an unidirectional for the other. A small speaker was located beside the computer screen which the students could have turned on and off. The professor used an umni directional microphone as his audio source.

Both locations, the classroom and the office, were connected by an audio video switch which helped us in simulating a remote teaching situation. In this scenario the professor instructed and supervise the students on a specific task; designing the set up for a multimedia teleteaching environment.

TASK DESCRIPTION

Using the global link the professor first provided an overview on the subject to all students. The students then instructed to work on the task individually while requesting an assistance

from the professor when necessary. The professor was instructed to "brows around" and to respond to the student questions. . Prior to the experiment, the professor and the students were given a short hands on training on the functionality of the technology.

We observed and analyzed the subjects' **Interaction, Task Exclusion and errors under** the following conditions :

Student - Prof.
Prof. -Student
Students -Prof.

Interaction

Different positions for the student's Camera and monitor front , shoulder, side, and combination.

Different sizes of Global screen and various distances from the students

Different positions and sizes of the Prof.'s screen

Different positions and distances of the Global camera

Interaction pattern With and without Timbuktu

Position of the Professor Camera CU, MS, LS

Room Architecture; classroom and office

Local sound level; pre set or controlled by the students

Global sound; pre set or controlled by the professor

MAIN FINDINGS

Global communication link was not used for interaction. It was only used to present information and to attract the Prof.. attention.

The camera and the monitor placed in a shoulder position was a very comfortable position in order to present documents but not for interaction.

Placing the camera and monitor in an upright position provided the most natural way for local communication but it wasn't suitable for the presentation of documentation

A medium shoot image of the professor was the most effective one for local and global interaction. Close up was too intense and long shoot revealed too many details that distracted the student.

The individual learning style of each students effected the way they executed the task using the technology. As such, the application should be flexible enough to be tailored to need of individual students.

The unidirectional microphone was far more effective than the omnidirectional one.

A shred workspace functionality is essential the teleteaching interaction and need to be investigated

When in local communication, a small window with the view of all classroom is required (PIP).

The transition between the global and the local mode was too slow. Transition time need to be decreased.

The Profs' monitor screen size was too small and it created difficulties in focusing on specific people.

Close up view of the student was the best for the local communication link.

Providing two local views shoulder and face to face simultaneously can be very interesting, yet, it might not be feasible from technological point of view and need to be investigated.

A feedback mechanism that reflect his image in the classroom was required by the Prof.

Based on the results of the experiment with the first prototype we proposed the following modification in the prototype and the scenario:

Prof. Office

Shared Workspace

Use of white board by the professor for global interaction.

Task exclusion
Novice vs. Expert use;The use of the system by a novice Prof.

Interaction

Use of the PIP technology by the professor to view the whole classroom while in local communication link.

Classroom

Disturbance

USE of Earphone by the students for a local communication link.

Interaction/ Task Excision

Use of two cameras- one for document presentation and one for local interaction. Is mixing between the two cameras possible?

Audio Disturbance - What is the level of the local and the global sound disturbance when students do not use the earphones

SECOND PROTOTYPE

Description
A second prototype was developed by us using the same technology as in the first experiment with some variations based on the suggestion from the first prototype trials We introduced two new components; the whiteboard and earphones. We also decreased the transition time between the local and the global communication link.

Three new subjects participated in this trial ; a professor and two students. The task was similar however this time we increased its difficulty by asking the students to design the computer configuration for the teleteaching setup. The students differed in their background and learning style; one was familiar with the technology and the other was not. The professor as well was novice to the use of the teleteaching application

We followed the same procedures as for the first prototype trial.

MAIN FINDINGS AND AREAS FOR INVESTIGATIONS

Training - The Prof. should be given a special training the professor on the use of the technology and its functionality . The professor then will deliver this information to the student at the beginning of a new teleteaching class and will also conduct with them a short hands on session using the application.

UI - The UI must support the Prof.. student local communication link. We suggested the use of an earphone and an Icon for signaling a request for a local communication link by a student to the professor.

AUDIO -When the global sound level is set higher, a larsen effect is created. We suggest suggested the use of earphones also for the global communication link. We also suggested that the global sound level would be set in advance and it would be controlled by the professor only.

AUDIO - The use of the earphones is suggested for the local communication link and need to be tested for these functions. Other options to overcome the larsen effect should be tested by the lab as well

Field of reference- Can the professor move while using the white board or does he need to sit and stay at the same position through out the course? Do we need a special camera to follow his movement?

Room Architecture - Is a U shape room architecture would have any effect on the Prof./ student interaction? Would it have any effect on the students/students interaction?

Two Cameras - Do the students need a document camera in order to present visual information? can we mix between the cameras signal? Can we use a mirror behind each student? If we change the set up to U shape can we use a white board behind each student to present the information (local white board)?

Shared workspace - How can the professor present his terminal 's screen to the students? Is using a global computer screen a good solution? What is the best software for local and global screen sharing? Which one is the best to run on an HP station

THIRD PROTOTYPE and SCENARIO

Description

Based on the results of our studies with the second prototype we modified the application and developed a third prototype. Our goal was to test the issue that were raised during the second trial. We added two new components to the setup; a whiteboard behind one of the student and a global shared workspace using a Barco in front of the classroom.

Three subjects participated in this trial ; a professor and two students. The professor participated in the first trial and was familiar with the technology. Both students were experts in using the technology.

The students sat in a row and were given a similar task to that of the second trial. They were asked to design the computer configuration for the existing setup. One student used Timbuktu as his shared work software, and the other used the whiteboard behind him for the same function. During the first half of the trial, the students were asked to use the earphones when engaged in a local communication link . During the second half the use of the earphones was optional.

Prior to the actual lesson we conducted a short training session with the Prof. He was given a 15min. introduction to the application and its functions. He then transferred this knowledge to the students and conducted a short hands on session.

Methodology

We used the same procedure as in the previous experiments.

MAIN FINDINGS and AREAS For INVESTIGATION

Task execution

The professor preferred using the white board behind him to present information formation to the classroom rather than using the global computer screen.

SCREEN SIZE - Although we replace the Prof. monitor screen to a bigger size it was still too small especially when he was engaged in a global communication link. When we used a wide angle lens for the global view of the classroom the situation got worse. To overcome the

problem the use of PIP technology was suggested whereby we split the screen and we show different parts of the classroom simultaneously. The danger here is that we can loose the spatial coherency and the natural motion of the classroom.

Audio Video Link- When a local communication link was in progress, seeing the Prof. image on the global screen without hearing him disturbed the other student who was not engaged in the conversation. We proposed three solutions to this problem:

1. Using a freeze image of the Prof., without motion

2. Portraying only a side view of the Prof. on the global screen.

3. "Background noise feedback"- All students will be able to hear (not to listen to) the local conversation as a background noise. Yet, the issue of privacy should be investigate here.

4. A message would appear on the global screen which will indicate "Prof. is engaged in a local communication link"

Audio - Both students were satisfied with the sound quality of the earphones. When instructed not to use them the sound from the loudspeaker was not a problem.

Training- The initial training had a positive effect on the use of the technology by the students and the professor. for example, the transition between the global communication link to the local one and vise versa was very smooth for both parties.

UI/ Gesture- Both students were concerned with the fact " how should I attract the Prof. attention? " How would I know that he received or saw my request? "Shall I use hands' gesture to attract the Prof. attention via the global screen". Our suggestion for this was that in to request a local communication link with the Prof. the students can use hands gesture yet, functions on the UI will be far more effective**.**

While using the UI to request a link a message should appear on the students' screen which will indicate that the message was received by the Prof. and that he will connect with the him ASAP. At the same time, a special earcon should indicate to the Prof. that someone is waiting for a local link.

Audio- An earcone should be heard while the Prof. is browsing around to indicate his entry to the student local screen.

Shared Work Space- The white board is a good solution only when a shred software is not available

Task- The task was too simple and not well prepared by the Prof. and might had an effect on the students' and the professor 's performance and use of technology. A well prepared challenging task should be taught by the Prof. The student(s) Prof. action and interaction should be tested under this condition.

Architecture- No special issue were raised regarding the classroom architecture; a row shape. A U shape room architecture should be tested as well.

*Illumination-T*he lighting conditions of the rooms should be tested as well especially when using the Barco.

FOURTH TRIAL

Description

To investigate the above we modified the existing prototype based on the result of the previous trial. This time the students were sited in a U shape position. Each student was provided with a whiteboard right behind him. The workspace of each student was illuminated by a local light .

Three subjects participated in this trial ; a professor and two students. All were experts with the use of the teleteaching application.

The students sat in a U shape classroom and were taught a new task: introduction to the use of Hyper Card. Both students were unfamiliar to the subject area and were eager learn it. Each student was provided with a white board behind him to present visual information. In addition, Timbuktu, the shared software was available to one of the students. The students were asked to use the earphones throughout the trial.

The Prof. who was an expert in the subject area was given one hour to prepare for this teleteaching course. Prior to the preparation we conducted a short training session with him. He then at the beginning of his class instructed the students on the various uses of the technology and conducted a short hands on trial.

MAIN FINDINGS and AREAS FOR INVESTIGATION

Training - During the training the students were not instructed on the use of the board and as a result did not use it. Teleteaching training should include all components of the teleteacing environment.

Illumination - The illumination of the student working environment was very bad. Using a table lamp effected the global view, however, it did not effect the local one. We had to find a solution to illuminate the classroom without effecting the Barco presentation.

Architecture- The U shape was as effective as the row one. The room architecture might have a greater effect on the interaction when a large number of students are involved.

Learning curve - The experienced Prof. and the students made very smooth transition between the global screen to the local one and vise versa.

Based on the five trials and the four prototypes we came up with the spec.[5] for the teleteaching set up. The flexibility of the application allowed us to conduct iterative modification during the implementation period.

Parallel to the implementation phase we conducted two studies which investigated the use of ISDN and ATM links for a teleteaching purposes. The results of which are presented in two

reports: Revital Marom, Lydia Goldberg, Pascal Gros, "Remote Technology Project Evaluation: An Assessment of a Teleteaching Environment" Research Report N° 93-003 Eurecom, December 1993. The second research report is still in its stage of analysis.

In the following section we will provide you with the a brief review of the two studies.

During the summer of 1993 we conducted an evaluation study of a real life teleteaching course. The two teleteaching sessions were conducted between Canada and France using an ISDN link switch (56kb/s). Based on qualitative evaluation methods, this study was designed to investigate which factors impact the learning and the teaching experiences in these setups from a technological and psychological point of view (e.g. social presence, technical capabilities of the system etc.). Thirty of Eurècom students participated in an HCI course that was taught by a professor situated in Canada. A computer scientist, a communications specialist and a psychologist were involved in the design and evaluation of this set up and application. The study revealed that the main problem of the teleteaching process is the lack of interaction between the professor and the students. This is a result of the inadequacy of the technology to support this function and the lack of experience by all participants in using the technology. One of our main conclusion is that To replace face-to-face instruction, an optimization of verbal and non-verbal interaction must be designed. Courses must be planned for a maximum information transference and acceptable levels of presence must be experienced by all participants. Perhaps a re conceptualization of what distance teaching is, would be necessary rather than attempting to duplicate a normal classroom situation. We should begin to think of new ways video transmitted data and images can be used for distance instruction. For instance, the system could be ideal for one-to-one tutoring between a student and a teacher where the shared workspace is used in conjunction with the audio/ video link.

In another evaluation study which was conducted during this fall an Image Analysis Coding course was taught simultaneously by a professor in France to two groups of students in two different locations. One group was copresent with the professor in Eurecom, France and the other was remotely located in EPFL, Switzerland. Our objective in this study was to conduct a comparison between a wide bandwidth link (ATM- 34 Mbps per second) and a narrow bandwidth one (ISDN 2x64) used for the teleteaching propose. The use of the two types of links allowed us to have two views of each site; direct view and peripheral view. The direct view served as a link between the students at the remote site and the professor and the peripheral one served as a link between the copresent group and the remote one. Both setups designed by the students who participated in the course. An iterative design modification was conducted by all the students at the end of each session.

Based on qualitative evaluation methods the study was designed to investigated some of the following variables which are specifically connected to remote teaching:

1. Professor orientation to the spatial configuration

2. Students orientation towards the special configuration

3. Use of artifact; camera, documents etc. by all participants.

4. Adaptation to the technology by all participants

5. Style / pattern of interaction by copresent and remote participants.

At present the data from this study is being analyzed in collaboration with a group of psychologists and sociologist from Xerox Parc and CNRS in Lyon.

The following are some of the preliminary unprocessed findings

Lausanne

The direct connection during the first day was via the ISDN link while the peripheral one was via the ATM. Consequently, the students found the setup to be frustrating and inadequate. The Professor image was often fuzzy and text could not be read clearly.

The students experienced lack of spatial representation of the Eurecom environment due to the location of the camera and the screens at Eurécom. They couldn't locate the direction of the gaze of their peers and the professor.

The delay in the image and sound transmission did not interfere with the learning process when no interaction was required. However, when needed to interact with the remote site "delays give the feeling that the other one is not reacting normally"

The remote students experienced difficulties in reading the text from the screen. It was suggested that a hard copy of the slides should be available to all remote students while the professor image will be projected onto the screen " The propose of a course is to listen to a professor explains verbally otherwise you can read the text on your own".

After changing the link so that the direct connection was via an ATM and the peripheral one was via an ISDN, the students reported that "a better quality picture of the professor made it more suitable for concentration during the lesson.

Synchronizing the cameras positions in both sites resulted in " a new feeling of a coherent view what we saw on the left side was on the left side in Sophia and vise versa"

Sophia Antipolis

The peripheral back view of the students in Laussane was perceived as " seeing the students from the back was not that interesting".

Using the ISDN link for peripheral awareness was perceived as "The quality of the video was not a problem for us it was sufficient enough to be aware who is there and what they are doing even if we had to make some efforts with the PicTel link.

During the first session, the Professor placed more emphasis on the remote class and neglected the local one. He was preoccupied with the quality of the images in Lausanne and was concerned with the increase learning difficulties for the EPFL group as a result of the distance.

When using the BETEL link for a peripheral awareness , the high quality peripheral image of the remote site supported the interaction between both groups of students. Yet the students report that "in teleteaching we should support the between the professor and the remote site and not between the students themselves"

During the first class the "peripheral" monitor was in a more dominant position and attracted more attention than the professor himself. This situation was changed when we placed the monitor in a less dominant place.

Appendex B: The use of adaptive filtering for echo cancellation

Reference [8] describes the principles of adaptive filtering based on the least-squares approximation, and develops a few algorithms, among which the Gradient, or Least Mean Square algorithm (LMS). The same notations as [8] have been adopted here. In this notation,

x(k) represents the sample at instant k
$\vec{x}(k)$ represents the vector of the N most recent samples at time k
$\vec{g}(k)$ is the vector of the N filter coefficients at time k

$$\vec{x}(k) = \begin{bmatrix} x(k) \\ x(k-1) \\ \dots \\ x(k-N+1) \end{bmatrix} \qquad \vec{g}(k) = \begin{bmatrix} g_0(k) \\ g_1(k) \\ \dots \\ g_{N-1}(k) \end{bmatrix}$$

where

y(k) is the sample coming from the microphone at time K
e(k) is the error at time k

The loudspeaker, room and B-microphone (see Figure 1) can be modeled by a system whose impulse response at instant k is $\vec{h}(k)$:

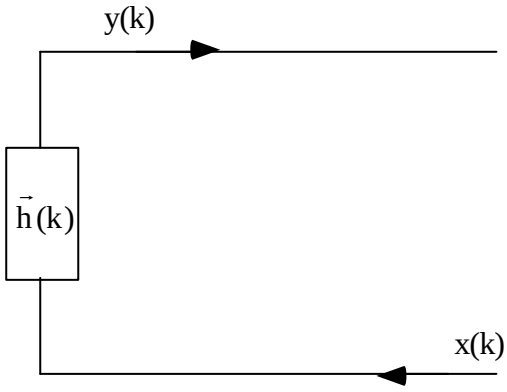

Fig. 1. Modelisation of loudspeaker, room and B microphone and where x(k) is the signal going to the speaker and y(k) the one coming from the microphone

To cancel the acoustical echo, the signal $\vec{g}^{T}(k)\cdot\vec{x}(k)$, with $\vec{g}(k)\equiv\vec{h}(k)$, must be subtracted from y(k) (see Figure 2). The filter $\vec{g}(k)$ then modelises $\vec{h}(k)$. In this case, the output signal e(k) is identically zero.

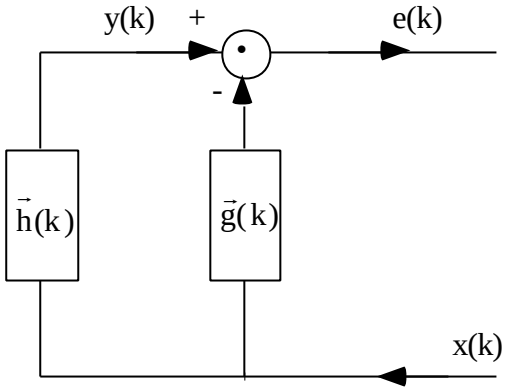

Fig. 2. There is no echo when $\vec{g}(k) \circ \vec{h}(k)$

However, as the system response $\vec{h}(k)$ (room response) is of infinite length, it is impossible to completely cancel echo with a finite-length filter $\vec{g}(k)$. Furthermore, $\vec{h}(k)$ may vary with time (for example by moving the microphone). The latter justifies the use of adaptive filtering, to make $\vec{g}(k)$ as close as possible to $\vec{h}(k)$, at any time. A last argument in favor of adaptive filtering is that there is no a priori knowledge of the room impulse response. Adaptive filtering is made in two distinct steps :

- the filtering, where the error is computed :

$$e(k+1) = y(k+1) - \vec{g}^{T}(k)\cdot\vec{x}(k+1) \qquad (1)$$

- the updating of the coefficients $\vec{g}(k)$, using the error e(k+1).

The second phase varies in complexity and in computing time, depending on the algorithm selected. For most real-time applications, the gradient (LMS) algorithm is preferred, as it has the advantages of an easy implementation, and shorter compute time. On a digital signal processor, filtering uses N cycles, and coefficient updating uses 2N cycles. For this algorithms, the updating takes the form :

$$\vec{g}(k+1) = \vec{g}(k) + Ke(k+1)\vec{x}(k+1) \qquad (2)$$

where K represents the adaptation step.

The two steps of adaptive filtering can be seen on figure 3.

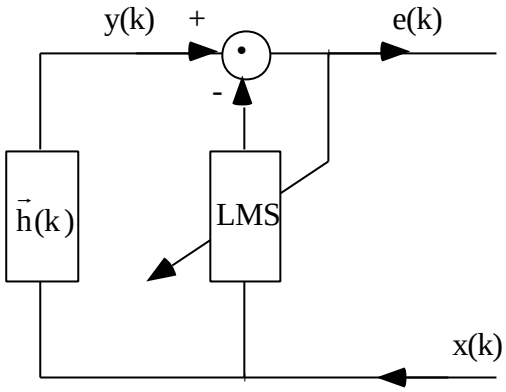

Fig. 3. Feedback of e(k) on the filter coefficients with emphasis on the adaptation process

The convergence time and stability of the system depend heavily on the value of the adaptation step K. If K is chosen too small, the adaptation will be very slow, but the error e(k), after convergence, will be very low and no instability hazard will be observed. If K is too high, not only the error may not be minimum, but there is a risk that the filter will diverge. In [8], one can see that the maximum value of the adaptation step depends on the length N of the filter and on the input signal power $\sigma_x^2(k)$, as follows:

$$K \leq \frac{2}{N\sigma_x^2(k)} \tag{3}$$

This inequality yields the following remarks:

• When N increases, K decreases, therefore the algorithm converges more slowly. This introduces a necessary trade-off between a correct echo cancellation ($N \to \infty$) and a fast adaptation.

• This stability relationship is approximately valid when it is close to adaptation process. It is therefore necessary to keep a security margin on K; generally, a factor of 4 is used.

• K may be a constant; in this case, it is necessary to take into account the least favorable case in the estimation of the signal power of $\vec{x}(k)$, to guarantee stability. However, this method would heavily penalize the adaptation speed for signals which do not use the full dynamics of the system. But one could be tempted to say the following :

> The convergence is a transitory phenomenon lasting only a few seconds, which can be neglected in regard to the duration of a teleconference talk. We can therefore use a small adaptation step K, accepting this penalty on the convergence, which simplifies the algorithm, and also improves echo cancellation.

The echo attenuation is improved by lowering K only if the computing accuracy is infinite, which is of course impossible. Consequently, the algorithm will converge only up to the point where the product Ke(k+1) is equal to 1 LSB. Beyond that point, the coefficient updating

reduces to $\vec{g}(k+1) = \vec{g}(k) + 0\vec{x}(k+1)$; there is no more adaptation. So it can be noticed that if K is small, the residual error will remain high, which proscribes this method in our case.

Taking in account the second remark above, the adaptation step becomes :

$$K = \frac{K_0}{N\sigma_x^2(k)} \text{ with } K_0 = \frac{1}{2} \qquad (4)$$

The signal power $\sigma_x^2(k)$ is then :

$$\sigma_x^2(k) = \sum_{i=k}^{k-N+1} x^2(i) = \vec{x}^T(k) \cdot \vec{x}(k) \qquad (5)$$

Calculating the signal power using this equation, is favorably heavy, as it needs N cycles on a DSP, as much as the filtering operation. It must be noticed that the accuracy of the result need not be high; a simple estimation may be more than enough. We will therefore use the following recursive estimation, using much less computing power :

$$\hat{\sigma}_x^2(k) = \beta\hat{\sigma}_x^2(k-1) + (1-\beta)x^2(k) \qquad (6)$$

where β represents the memory factor. If $\beta = 0$, the estimator has no memory, and in this case, $\hat{\sigma}_x^2(k) = x^2(k)$, which no longer represents the power of the $\vec{x}(k)$ vector, but the power of the sample x(k) would have disastrous consequences on the stability of the filter. Actually, the estimator must have enough memory to "remember" the contribution of a new sample until the vector $\vec{x}(k)$. has been computed. Taking the value

$$\beta = \sqrt[N]{0.5} \qquad (7)$$

the estimator $\hat{\sigma}_x^2(k)$ will still contain a contribution of 50% from the oldest sample x(k-N+1), which is enough if one is aware of the security margin taken in account in the computation of K (factor K_0).

There is another recursive method for the computing of signal power :

$$\sigma_x^2(k) = \sum_{i=k}^{k-N+1} x^2(i) = \sigma_x^2(k-1) + x^2(k) - x^2(k-N+1) \qquad (8)$$

The contribution of the oldest sample is subtracted, then the newest is added. It can be noticed that this method is not an approximation. One must take in account that squaring a number needs twice the word width. If double precision is not conserved, an error will accumulate with every iteration, which will make $\sigma_x^2(k)$ grow to infinity.

We will analyze the three possible scenarios between both participants A and B, then examine how the adaptation evolve and the effects of the filter on speech.

• When neither participants is speaking

In this case, the filter has no effect on the signals, as these are null or very weak. However, the adaptation step tends to infinite because signal power is tending to zero. It must be noticed that theoretically, the system should not adapt; introducing $\vec{x}(k) = \vec{0}$ in (2), we obtain $\vec{g}(k+1) = \vec{g}(k)$. Nevertheless, as the signal $\vec{x}(k)$ is actually never zero, the system adapts on ambient noise. At first sight, this situation seems favorable, since the system is able to adapt in the absence of speech, but actually it introduces a problem, as we will see it in the last case. So provision must be made that the system stops adapting when the participants do not talk, or in other words, to saturate the adaptation step K is assigned to a value K_{max} to suppress or at least to reduce ambient noise.

• When one participant is talking

If participant A is talking, the canceller on B will adapt. On the other hand, on side A (supposing A also has an echo canceller), $\vec{x}(k)$ is zero, so the adaptation step will tend to infinite. Substituting $\vec{x}(k) \cong \vec{0}$ in (1), one can see that the error signal becomes equal to the output $e(k+1) \cong y(k+1)$. Introducing that in (2), we obtain $\vec{g}(k+1) \cong \vec{g}(k) + Ky(k+1)\vec{x}(k+1)$. Remembering that K is very high, and that $\vec{x}(k)$ is never equal to zero because of the ambient noise, the system will try to adapt, but randomly, which is definitely not desirable, and that will in turn distort the voice of the A talker.

• When both participants talk

First, it must be remarked that this happens normally only during a short while (except in French political debates); one will finally let the other speak, and we go back to the preceding scenario. During the short moment of overlap, the adaptations at A and B will change randomly and both voices will be distorted.

This shows the importance of saturating the adaptation step to K_{max} to avoid voice distortion and a bad echo cancellation. With n representing the RMS value of the noise in the signals x(k) and y(k), the following relation on K_{max} allows to stop the adaptation on the ambient noise :

$$K_{max} n^2 < 1 \text{ LSB} \Leftrightarrow K_{max} < \frac{1 \text{ LSB}}{n^2} \qquad (9)$$

This inequality is not very restrictive due to the squaring of the noise (which is supposed weak). On the other hand, the restriction on scenario 2 is more important :

$$K_{max} n \sqrt{\sigma_x^2} < 1 \text{ LSB} \Leftrightarrow K_{max} < \frac{1 \text{ LSB}}{n \sqrt{\sigma_x^2}} \qquad (10)$$

Consequently, only the latter expression will be considered. It is not easy to know at priori the RMS value of the room noise. Nevertheless, a good estimation of it can be made, supposing that the room is quiet and that the microphone is close to the participant. In this case, the signal/noise ratio SNR close to the microphone is estimated to about 45 dB. On the other hand, it is reasonable to estimate that the signal uses up almost the full dynamic range of the

system. In this case, for a speech signal, the RMS voltage of the input signal $\sqrt{\sigma_x^2}$ is about 0.25 (versus the full dynamic range). From this, we can deduce the noise voltage n:

$$\frac{\sqrt{\sigma_x^2}}{n} = S/B \Leftrightarrow n = \frac{\sqrt{\sigma_x^2}}{S/B} \Rightarrow n\sqrt{\sigma_x^2} = \frac{\sigma_x^2}{S/B} = 0.35 \cdot 10^{-3} \qquad (11)$$

With a 16 bit quantification, we find for K_{max} :

$$K_{max} < \frac{1\ LSB}{n\sqrt{\sigma_x^2}} \cong 0.87 \qquad (12)$$

Algorithm

The different steps described are summarized :

Initialization

{

#define N=length filter

#define K_0=0.5 /* security factor */

$\vec{x}(0) = \vec{g}(0) = \vec{0}$

$\hat{\sigma}_x^2(0) = 0$

$\beta = \sqrt[N]{0.5}$

$\tilde{K} = \frac{K_0}{N}$

$K_{max} = 0.87$ for 16 bits quantification

}

For every instant k

{

read x(k) et y(k)

insert x(k) in $\vec{x}(k-1)$

compute $e(k) = y(k) - \vec{g}^T(k-1) \cdot \vec{x}(k)$

compute $\hat{\sigma}_x^2(k) = \beta\hat{\sigma}_x^2(k-1) + (1-\beta)x^2(k)$

compute $K = \frac{\tilde{K}}{\hat{\sigma}_x^2(k)}$

if $K > K_{max}$ then $K = K_{max}$

update $\vec{g}(k) = \vec{g}(k-1) + Ke(k)\vec{x}(k)$

send e(k)

}

Computing time and performances of the system

Computing time varies linearly with length N of filter. Filtering and coefficient updating require respectively N and 2N instructions, and a few more instructions are necessary for the other steps of the algorithm. It is important to notice that these computing times are valid only

considering a signal processor able to execute in parallel and in one clock cycle all the following groups of instructions:

- multiplying two registers followed by an accumulation,
- loading a coefficient and a sample in registers with auto-incrementation of pointers,
- decrementing a counter with a conditional jump at the beginning of the loop.

In this application, a telephony pass band has been used (300-3400 Hz), allowing a sampling frequency of 8 kHz. With a DSP clock frequency of 20 MHz, we have 2500 machine cycles at our disposal, distributed as follows:

- 820 cycles for filtering,
- 1640 cycles for coefficients update,
- 40 cycles for computing the adaptation step and other routines.

A filter length of 820 with a sampling frequency of 8 kHz yields a time span of 102.5 ms. This time represents a path of about 35 m for sound propagating in air. Consequently, the canceller cannot "see" any echo path longer than 35 m, as it will not be able to find a correlation between loudspeaker output and the microphone input signals. This emphasizes the importance of the room reverberation time as it allows to estimate the residual level 102.5 ms after the extinction of the source.

When evaluating the prototype, we measured the echo attenuation in a test room at 20 dB. A measure of the room reverberation time showed us that the signal decreased by 20 dB in 102.5 ms. The performances of the system seem to be closely related to the acoustics of the particular room.

Appendex C. Measured Performance of the teleteaching platform

Measurement tool and scenarios

A simple tool, called ttcp, was used to measure throughputs. This tool allows users to create messages of various lengths for studying memory to memory transfer effects between two workstations. It is necessary to send a large enough amount of data lasting several seconds to obtain accurate measurements. In this study, the averaged memory to memory throughputs of TCP/IP and UDP/IP were gathered, by varying the messages sizes, sending the same messages repeatedly until 16 MB of data were transferred, and repeating each measurement ten times. The default parameters in Table 1 were used.

Parameters in Bytes	TCP/IP	UDP/IP
Socket Send Buffer Size (SO_SNDBUF)	24578	9000
Socket Receive Buffer Size (SO_RCVBUF)	24578	18032
Window Size	24578	N/A
Maximum Segment Size (MSS)	4312	N/A

Maximum Transfer Unit (MTU)	4352	4352
Total Amount of Data Transferred	16M	16M
Number of Trails	10	10

Table 1 Default parameters used in this study

The Round Trip Time (RTT) measurements are made using a standard UNIX utility called ping. It allows users from any host to send a small packet to a remote host and wait for the returned packet which contains the RTT information, and this operation can repeat a number of times (100 is used in this study) and statistics on RTT can then be collected.

During the measurement, both CPUs were dedicated for transmitting or receiving data with no other active processes running while the measurements were taking place, and there was no other traffic in the FDDI ring and at the BETEL links between EPFL and EURECOM when measuring BETEL teleteaching network performance. The checksums were on all protocols except UDP.

Measurement in local FDDI environment

The measurements were taken between two SUN Sparc10 stations model 51 running Sun OS 4.1.3, with a clock frequency of 40 MHz and 64 MB of RAM. They were placed one meter apart and connected by an FDDI ring at the Laboratory of Telecommunications at EPFL. The FDDI interfaces used were SunLink FDDI/S (version 1.0) with a Maximum Transfer Unit (MTU) of 4352 bytes.

Figure 1 shows that TCP/IP and UDP/IP throughputs vary with message sizes. For larger messages, UDP/IP has larger throughputs than TCP/IP while the converse is true for small messages. This is because smaller messages have more overhead in UDP.

Moreover, there is a drop in UDP/IP throughput when the message size is near 4325 or 8649 bytes. This is the fragmentation effect. The UDP/IP throughputs are reduced because fragmentation and reassembly need additional processing and CPU power. Moreover, the throughputs decrease since the first fragmentation takes place. If one of the fragments is lost, the UDP datagram cannot be reassembled and so it is discarded. Hence more packets are lost for smaller throughputs.

Furthermore, the percentage of the UDP packet that are lost also increases with message sizes. The fragmentation effects discussed above is one of the reasons. Another is the buffer overrun at the receiver side. The sender transmits much too fast for the receiver to process all incoming packets. The larger messages occupy more buffer space. Therefore, the buffer overrun at the receiver end takes place more frequently for larger messages. In addition, about 10% of the UDP/IP messages are lost with message size of 128 bytes, due to high processing overheads and high frequency of I/O interrupts for very small messages.

On the other hand, the TCP/IP throughputs do not perform fragmentation because the MSS is less than the MTU size. The throughput increases steadily until it reaches its maximum at about

5 MB/s. For transporting the same amount of data, larger messages need less system calls and processing overhead, hence yield higher throughputs.

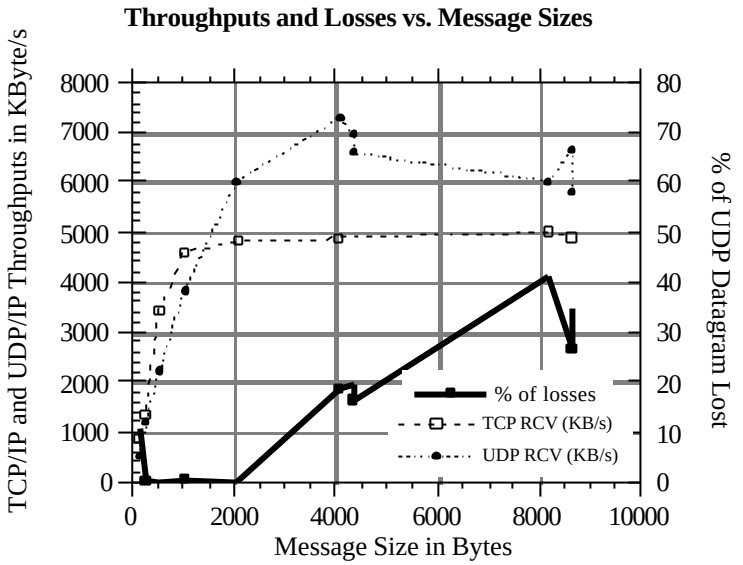

Fig. 1. Performance measurements between two Sparc 10 stations over FDDI

In summary, this study using ttcp to measure maximum throughputs suggests that UDP/IP (without UDP checksum) could achieve maximum throughputs of 7.3 MB/s for message size of up to 4K bytes with maximum 20% losses while TCP/IP has a maximum throughput of 5 MB/s for message sizes over 4K bytes. Thus, if the MTU of 4K is used between EPFL and EURECOM, UDP/IP could have a maximum of 6 MB/s uni-directional video transmission using 2K byte video packets with maximum 0.3% losses; 256 byte UDP/IP datagrams could be used if maximum 0.4% of loss rate is acceptable for audio transmission.

Measurement on the BETEL teleteaching platform

The BETEL teleteaching network performance measurements were taken between two Sun Sparcstations 10 running Sun OS 4.1.3. The Sparcstation at EPFL, equipped with SunLink FDDI/S LAN interface, is of model 51 with a clock frequency of 40 MHz, while the another one at EURECOM is of model 31 (36 MHz clock frequency) with Daul-Attach FDDI interface from Network Peripherals).

The maximum TCP/IP throughput measured, using the same tool and same method as the previous study (described in section 5.1.1), is shown in Figure 2. At steady state, about 1.05 MByte/s bandwidth on top of TCP/IP level can be available for the teletutoring application on the BETEL teleteaching platform. Since the two workstations do not have exactly the same configuration, using Sparc10 model 51 as the sender and Sparc10 model 31 being the receiver can obtain higher TCP/IP throughput (1.1 MByte/s when TSDU is 2 KBytes or greater).

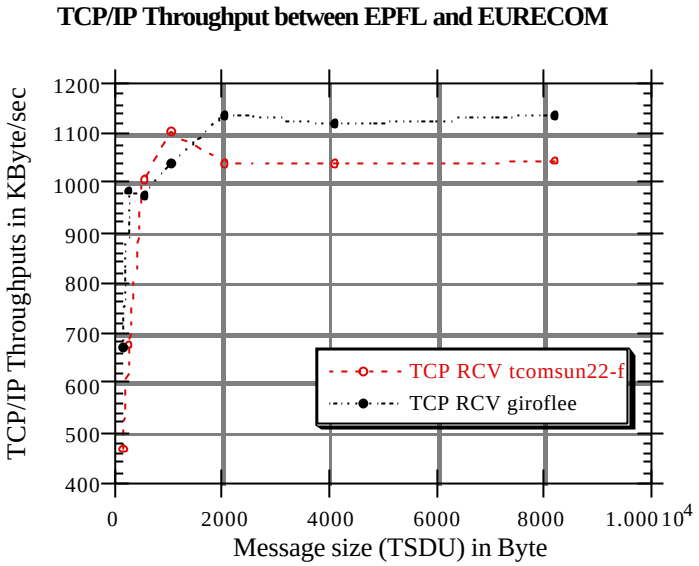

Fig. 2. Maximum TCP/IP throughputs in the BETEL teleteaching platform

The average RTT needed to travel between the two Sparc10 stations via the BETEL teleteaching network is given in Figure 3. The measurement is done using ping utility. This graph shows that it will take on average about 13 ms for a small TSDU of 128 bytes to make a round trip, for example, from EPFL to EURECOM, and back to EPFL.

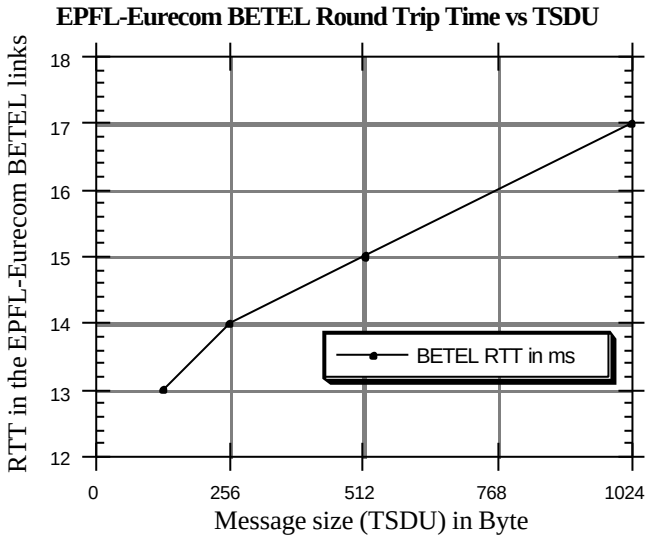

Fig. 3. Average RTT in BETEL teleteaching platform

These results indicate there is enough bandwidth available for the teleteaching application, and confirm that the BETEL teleteaching network is a long fat network which has large bandwidth-delay product. TCP is not suited for transport real-time audio and video data and the alternative is thus UDP. Moreover, the bottleneck here is not in the network (i.e. bandwidth) but at the endsystems.

Appendex D. Theoretical Workstation Performance Evaluation

1. Introduction

The goal of this document is to evaluate the performance limitations of the workstation to be used in the Betel project, especially with regards to the video conference part of the application. A basic assumption is that only two-party video conferences will have to be established.

The system under study comprises:

- a Sun SPARCstation 10 model 51 (50 Mhz, 64 MB RAM, 1 GB Disk)
- a Parallax Xvideo acquisition board including the CL550B JPEG compression chip from C-Cube Microsystems
- an FDDI board (Sunlink FDDI/S)

The document is partly based on similar studies of other systems using the same compression chip and a comparable workstation architecture [12, 13]. This study will focus on three different topics, namely the performance limitations introduced by the Parallax board, by the system bus and by the network and will try to deduce from these evaluations theoretical upper bound values for image size, number of images per second as well as an estimation of the maximum end-to-end throughput.

2. Performance of the video acquisition board

2.1. Architecture of the Parallax board

The architecture of the Parallax video acquisition board is presented in figure 1. The analog video signal is digitized and the resulting data is stored into the card's own frame buffer. The frames are then transferred via the video channels to the C-Cube chip to be compressed. Finally the compressed signal is sent via DMA (Direct Memory Access) to the workstation's main memory.

2.2. Performance evaluation

According to the previously mentioned architecture, two elements are likely to impose a limitation to the performance of the board, namely the compression chip and the pixel bus.

According to the compression chip specification, the chip is able to process 30 images per second at full resolution, and should not therefore, if these values are confirmed, be a limitation.

On the other hand, the throughput at which the video frames are moved on the board itself are not unlimited. As a matter of fact, the pixel bus of the Parallax board comprises two 15 Mpix/s video channels - one unidirectional, one bi-directional - capable of carrying 45 Mbytes/s of data [7]. That is, one pixel can be processed every 66.67 ns.

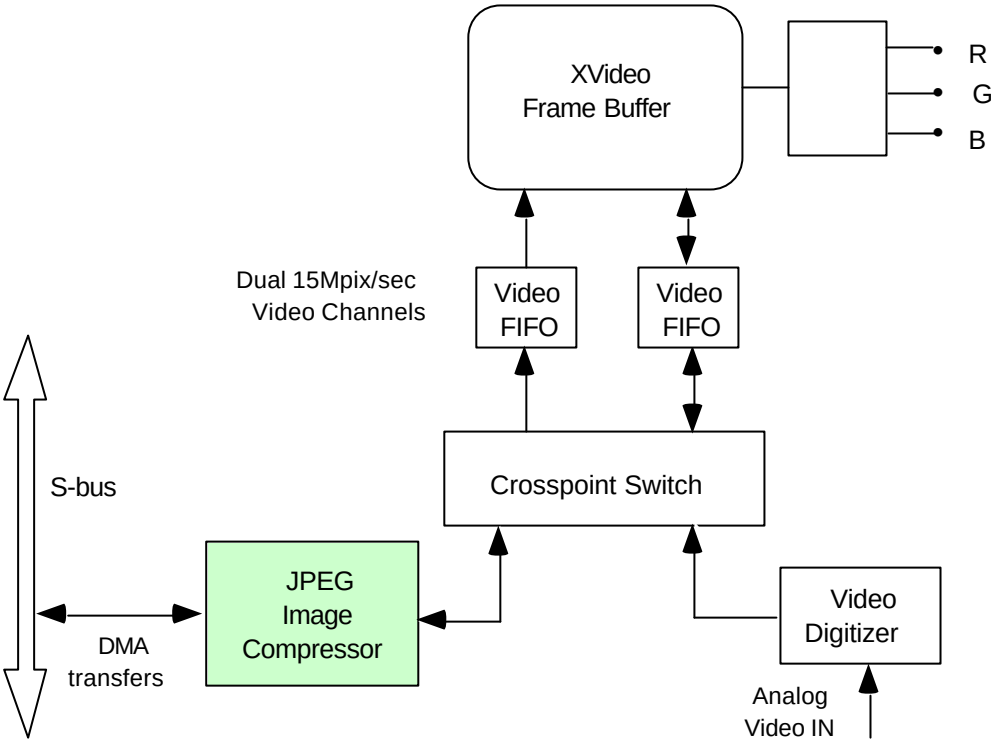

Figure 1: Simplified Architecture of the Parallax board

The compression or decompression time is the same and is the line size in x (line_size) plus the line oriented overhead times the line count in y (line_cnt) plus the vertical oriented overhead. The turnaround time is the time it takes to change the mode on the CL550 from decompression to compression or vice versa. Its value is 700 write cycles at 240 ns per cycle, namely 168 µs.

In the context of a two-party video conference, the Parallax board has to compress the outgoing video signal and to decompress the incoming one. The full cycle of operations within the refresh time of one image is then:

(compression time) + (turnaround time) + (decompression time) + (turnaround time)

with

compression time for one image : $C_t = ((line\_size + 16)*(line\_cnt + 2))*66.67ns$ [2]

decompression time for one image : $D_t = ((line\_size + 16)*(line\_cnt + 2))*66.67ns$

[2] The line oriented and vertical oriented overhead values come from [1]

The real-time constraints of the video conference impose the following relation, where n is the refresh rate for the images, or in other words the number of images per second:

$$1/n >= C_t + D_t + 2*\text{turnaround\_time}$$

The two important parameters which can have an influence on the performance of the board are then the image size and the refresh rate. The above relations allow us, assuming some simplifications, to express the conditions these parameters must fulfill to allow real-time video processing.

Assumptions:

- the line oriented and vertical oriented overheads are neglected as a first approximation
- the turnaround time is neglected (for PAL images its value is about 100 times smaller than the compression time)
- length and width of the image in a 4:3 ratio

The maximal values for the image size and refresh rate are then expressed as follows:

$$1/n = 2*(\text{line\_size} * (3/4 * \text{line\_size}))*66.67*10^{-9}$$

$$\boxed{\text{line\_size}^2 * n = 10^7}$$

Examples: n = 25 images/s --> max size = 632 x 474

size = 768 x 625 (PAL) --> n_{max} = 16 images/s

The above expression gives a good approximation of the refresh rate that can be expected for a defined image size and vice versa. A more accurate calculation gives the following results, which imply a data rate (uncompressed video) of about 173 Mbps.

Image size	**Refresh rate**
620 x 465	25 images/s
768 x 625 (PAL)	15 images/s

The previous results show that the transmission of TV-like quality (PAL, 25 images/s) is beyond the capabilities of the Parallax board.

3. Performances of the workstation

3.1. Architecture of the SPARCstation

Thanks to the efficient architecture of the Parallax board, in particular the presence of the frame buffer and of the compression chip on the board itself, only compressed image frames are sent on the system bus. For each video connection, data has to cross the system bus twice (from the frame buffer to the system memory and from the system memory to the network interface), that is to say four times for a two-party video conference.

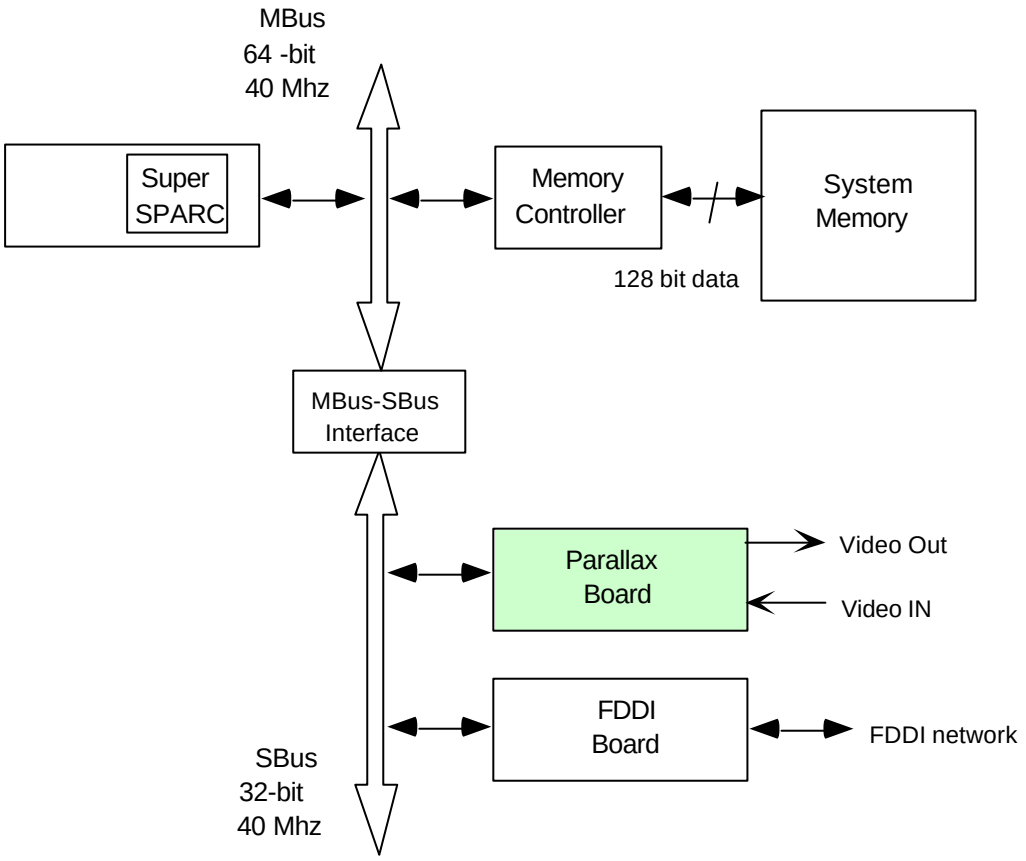

Figure 2: SPARCstation 10 Architecture

According to the architecture of the SPARCstation (figure 2), the bottleneck along the previously mentioned path is the SBus itself, as the MBus is twice faster [9]. Actually, an important benefit of this architecture is that even when the SBus is performing data transfers at peak transfers rates, there is still significant bandwidth available to the CPU on the MBus, e.g. to run the shared application.

If one assumes a sustained bit rate of 50 Mbytes/s on the SBus (according to Sun's technical paper [9]), as well as an optimal data multiplexing on the bus, the bandwidth available for each video stream is then 12.5 Mbytes/s. This value is, however, far too optimistic as it

neglects the overhead for bus arbitration and above all the behavior and load of the CPU itself. Thus a more accurate model has to be investigated.

3.2. Data path through the system

Several aspects of the workstation's architecture, apart from the processor, have an influence on the overall performance for networking, in particular the system bus and the memory subsystem characteristics. In order to better quantify their respective impact, a first step consists in examining the path taken by the data as it passes through a conventional protocol stack. This path is illustrated in figure 3 [14].

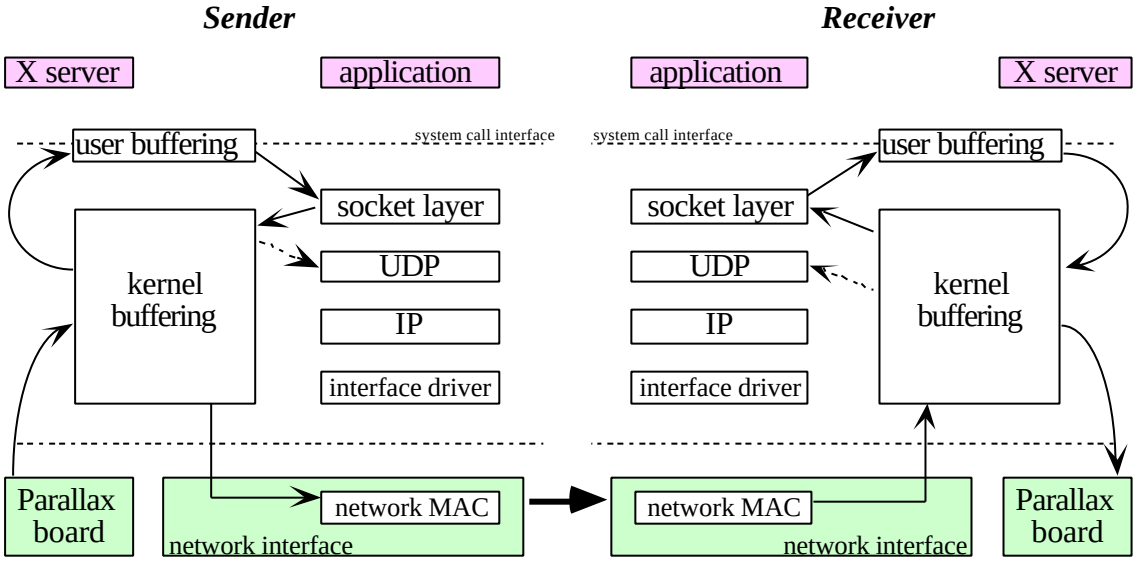

Figure 3: Data paths in a conventional protocol stack

In the Betel configuration, the data is first sent via DMA from the Parallax board to the system memory and then copied in the user memory to be processed by the application (namely the Audio Video Supervisor (AVS)). A supplementary copy operation, which does not appear explicitly on this figure, happens in the user buffer space, due to the software architecture of the Parallax board. As a matter of fact, the frames copied from system memory to user memory are first processed by the X server and then transferred to the X client (AVS). This latter invokes then a system call to send the data. The data is copied by the socket layer into the kernel memory ('mbufs' buffers), where it is processed by the protocol layers.

At that stage the UDP checksum is optionally computed, according to the corresponding setting. The deactivation of this checksum could avoid a supplementary and perhaps not necessary access to the data.

Finally the data is copied out to the network interface (FDDI board) using DMA. On the receive path the behavior of the system is similar, implying the same amount of data movements between the different subsystems of the workstation.

3.3. Performance model

The theoretical performance of the workstation may be evaluated by using a simple model presented in [10]. In this model the service time, namely the time taken by an individual server to process a packet, is broken up into two parts:

- A fixed per-packet service time, including:
  - filtering of packets by the network interface
  - datalink layer, network layer and transport layer processing
  - interrupt processing, memory management and context switching

- An incremental service time that varies with packet size, including:
  - data movement between host main memory and the network interface
  - data movement between host main memory and the video board
  - data movement from (to) system memory to (from) user memory
  - error checking overhead (optional)

Finally the throughput (number of packets per second) is defined as follows:

$$\text{Throughput} = \frac{1}{\text{fixed service time} + \text{incremental service time}}$$

3.4. Evaluation of the fixed service time

As far as the fixed service time is concerned, theoretical estimations and measurements for a comparable system (DECstation 5000/200) have evaluated its value at about 400 µs per packet [10]. In order to estimate this value for the SPARCstation a simple approach is to scale this service time with the CPU speed and SPECmarks rating of the system. According to the hardware specification of the two workstations (see the performance quick reference card in [16] p.41, for example), the performance ratio is approximately 2 in favor of Sun. Scaling by this factor, the fixed per-packet service time for the SPARCstation is expected to be about 200µs.

3.5. Evaluation of the incremental service time

The incremental processing overhead of packets is primarily due to data movement, namely:

- DMA video acquisition board <--> system memory
- copy system memory --> user memory
- copy user memory (X server) --> system memory (X client)
- copy user memory --> system memory
- DMA system memory <--> network interface (FDDI board)

- (read system memory --> UDP layer)

According to the measurements results published in [17], the sustained CPU/Memory bit rate of the SPARCstation is about 220 Mbps for copy operations and 350 Mbps for read operations. These relatively low values, at least in comparison with the 2300 Mbps theoretical peak bit rate, are due to the low cache hit rate typical in such data movements. Subsequently, the transfer of a packet from user to system memory (and vice versa) will take about 149 µs.[3]

Regarding the DMA transfers, it is generally possible to transfer large blocks of data in a single bus transaction, thereby achieving transfer rates close to the limits of the main memory and I/O bus speed. The data transfer can proceed concurrently with activity by the processor, although contention for main memory access may induce processor stalls during periods of heavy DMA traffic.

A simplified model of these transactions is to have two servers involved in processing a packet, one is the DMA engine and the other is the processor itself performing the fixed per-packet processing at the driver and higher layers. According to [10] the host CPU is the bottleneck in this model. Therefore these DMA transfers may be considered as happening in parallel with CPU normal packet processing and having almost no impact on the overall performance of the system. However, to take into account the contention for memory, in other words to quantify the effect of the DMA engine stealing cycles from the processor, one cycle is considered as being stolen from the CPU for every long word (4 bytes) transferred. For the 50 MHz CPU of the SPARCstation, this penalty implies a supplementary service time of about 20,5 µs per packet.[4]

3.6. Performance evaluation

Applying the previously mentioned values to the expression defining the throughput and assuming UDP packets of 4 Kbytes, the maximum transfer rate delivered by the SPARCstation for an unidirectional video connection may be evaluated, namely:

UDP without checksum:

$$\text{throughput} = \frac{4096}{200\ \mu s + 3.(145\ \mu s) + 2.(20{,}5)\ \mu s} = 6 \text{ Mbytes/s}$$

UDP with checksum

$$\text{throughput} = \frac{4096}{200\ \mu s + 3.(145\ \mu s) + 2.(20{,}5)\ \mu s + 94\ \mu s} = 5.3 \text{ Mbytes/s}$$

[3] transfer time $= \frac{4096 \text{ [bytes/packet]} \cdot 8 \text{ [bits/byte]}}{220.10^6 \text{ [bits/s]}} = 149 \cdot 10^{-6}$ [s/packet]

[4] Numerical results are for 4096 bytes packets

As far as bi-directional video connection is concerned, the half of the above values may be considered as an upper bound value of the throughput, implying a maximum data rate across the system of about 3 Mbytes/s (2.65 Mbytes/s with checksum).

The previous results, which imply a 100% CPU utilization, are consistent with those obtained in similar studies with roughly equivalent HP and DEC workstations [10, 16].

4. Performance of the network

Another possible performance bottleneck of the video conference system is given by the network (FDDI, ATM) and the access interface to the FDDI ring.

Regarding the bandwidth at disposal, the theoretical maximum values for a bi-directional video connection are the following:

- ATM network : 4,25 Mbytes/s [5]
- FDDI network : 6,25 Mbytes/s [6]
- FDDI interface : ~ 4,5 Mbytes/s [7]

The maximum data rate of the network is then within the same order as its corresponding value in the workstation. However, it is worth mentioning that due to the different protocol overheads necessary to transport IP packets over the ATM network, the effective throughput (i.e. of useful information) of this latter will not exceed 2,6 Mbytes/s. [17]

5. Comparison with experimental results

In order to evaluate the relevance of the previous estimations, the theoretical results obtained are compared with experimental measurements performed on the system under study. These studies have been done in a very local environment without taking into account the Betel network configuration, in particular the overheads arising from the terminal adapter and the routers. These studies are only concerned with the performance of the customers premises networks but can be used to validate the theoretical model presented in chapter 3. Two types of measurements have been performed so far,

- the first one addresses raw data transfer at the transport layer (Appendex C)
- the second one measures the end-to-end throughput of a video connection obtained with the current implementation of the Audio Video Supervisor.

---

[5] ATM link: 34 Mbits/s in each direction (maximum value)

[6] FDDI = shared media --> 50 Mbits/s for each direction

[7] According to Sun

5.1. Data transfer performance

The experiment considered in Appendex C addresses the maximum performance of transport layer protocols (in particular UDP) in function of the packet size. The data is transferred from workstation memory to workstation memory via an FDDI ring. Theoretical and experimental results are compared in the context of an unidirectional data transfer.

If one applies the theoretical model presented in chapter 3.3 to this raw data transfer, the incremental service time comprises the copy operation from user memory to system memory and the DMA transfer from system memory to the network interface (i.e. penalty imposed on the CPU by the DMA transfer). The throughput can then be expressed as follows:

$$\text{throughput} = \frac{p}{FST + IST(p)} \qquad (p = \text{packet size in bytes})$$

with $FST = 200$ [µs/packet]

$$IST(p) = p\left(\frac{8}{220 \cdot 10^{-6}} + \frac{1}{4 \cdot 50 \cdot 10^{-6}}\right) = 45 \cdot 10^{-9} \cdot p \text{ [s/packet]}$$

The experimental and theoretical curves are represented in figure 4.

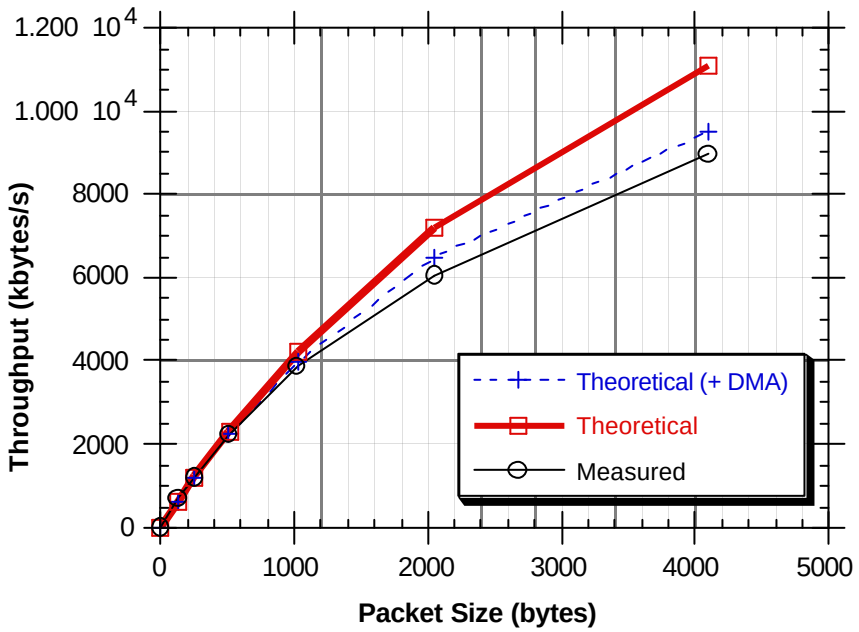

Figure 4 : Comparison of Data Transfer Performances

The throughput expressed in function of the packet size is an expression of the form

$$f(p) = \frac{p}{a + b.p}$$

Thus the slope at origin is equal to $1/a$, or in other words is the inverse of the fixed service time. The good match of the two curves for small packet sizes show that the estimated FST value of 200 µs is very accurate.

For large packets, however, the difference between theoretical and measured results becomes significant. This can be explained if one considers a maximum throughput of the FDDI interface of about 9 Mbytes/s, which for large packets tend to be the bottleneck in the system. Moreover, it is also possible that the effect of DMA transfers on the CPU (contention for memory) has been under-estimated. The dotted curve on the previous figure (which takes into account the DMA transfer time or in other words does not consider it as happening in parallel with normal CPU processing) tend to confirm this explanation.

5.2. Video Communication

The results presented in chapter 3.6 seem far too optimistic when compared with the maximum throughput measured on an unidirectional video connection established between two workstations. As a matter of fact, the mean end-to-end data rate measured is about 5,9 Mbps, in other words eight times smaller than its corresponding theoretical value.

This difference arises from the implicit assumption made in the performance evaluation that the different operations performed by the workstation to capture and transmit moving images (image digitization and compression, image transfer into host memory, packetization and data transfer) occur in parallel. In fact, this approach doesn't take into account the way the application controlling the video conference is implemented. In the system considered the application is implemented as a single process and thus all the operations described previously occur sequentially.

In addition, and according to a bug report from Parallax, the maximal performance of the board is lower than previously supposed, namely 13 PAL images per second for an unidirectional connection.

According to the previous remarks a more accurate performance evaluation may be performed, based on the following parameters:

- image size = 384 x 287 (the quarter of a PAL image)
- average compressed image size : 20 Kbytes (i.e. five 4 Kbytes UDP packets are needed to transfer one compressed image)
- assumed DMA speed of 50 Mbytes/s

To take into account the application overhead incurred during image processing (in particular context switching, X overhead), the fixed service time per packet has been doubled. To allow computation of the end-to-end throughput, the incremental service time which refers to images is expressed in terms of average processing time per packet.

The incremental service time required by each sequential operation can be estimated (according to chapter 3.5):

• image digitization and compression	=	19,2	ms/image	-->	3,8 ms/pkt
• DMA to system memory (SM)	=	409,6	µs/image	-->	82 µs/pkt
• copy SM - user memory (X server)	=	744,7	µs/image	-->	149 µs/pkt
• copy UM (X server)- UM (X client)	=	744,7	µs/image	-->	149 µs/pkt
• copy UM (X client) - SM				-->	149 µs/pkt
• DMA SM - network interface				-->	82 µs/pkt

The application of the previous values to the expression presented in chapter 3.3 gives the following result:

theoretical throughput = 6,75 Mbps

This result is very close to the measured value of 5,9 Mbps. The difference should mainly arise from the variable size of a compressed image which is in general not an integer multiple of the optimal packet size and thus reduces the transmission efficiency.

6. Conclusion

This study has tried to explore the potential performance limitations of the video conference system to be used in the Betel project.

Regarding video acquisition, theoretical calculations have shown that PAL images (768 x 625) can be refreshed up to 15 images per second, whereas the maximum size at 25 images/s is about 620 x 465. These values arise from the 15 Mpix/s maximal throughput between the frame buffer and the compression chip but seem somewhat optimistic in comparison with the actual performance published by the manufacturer in a recent bug report.

As far as the throughput is concerned, this document has shown that the bottleneck seems to be the workstation itself, and in particular the video acquisition board. Sustained data rate of about 6 Mbytes/s (unidirectional) seems to be the upper bound limit of currently available hardware. Measurements of the effective throughput reached by current implementation of the video conference system shows a drop in performance by a factor 8 in comparison with the expected value. A part of this difference could be suppressed, however, by better exploiting the inherent parallelism of the different operations involved in capturing and transmitting video. Not allowed by the time frame of the Betel project, the conception and implementation of a more efficient video control software could be a matter of further study.

In conclusion, even if it is not optimal for supporting video conference, the hardware architecture of the SPARCstation seems to be efficient enough to support the teleteaching application foreseen in the Betel project. In particular, the CPU should be powerful enough and should have enough bandwidth free on the MBus to perform other activities in parallel with the video conference. However, as the potential of the workstation is not totally exploited, the goal of fully utilizing the 34 Mbps links at disposal cannot be achieved in the

short term if only one workstation is allocated to each participant. Therefore the use of several workstations per participant, each one fulfilling a specific task (e.g. capture and send video, receive and display video, share teaching application), seems the safest short term solution towards this goal.

References

[1] Martin, O. H., "Broadband Exchange over Trans-European Links (BETEL)", Proc. SMDS Conference, Amsterdam, November 1993.

[2] Martin, O. H., "A Perspective on the shift/BETEL Project", the 15th Speedup Workshop on "Visualization and Networking", Lugano, Switzerland, on March 17-18, 1994.

[3] Y. Le Moan, "Data Transfer Service Specification", BETEL internal document, CIT-5, April 1993.

[4] Y. Le Moan, "Traffic Matrix", BETEL internal document CIT-8, May 1993.

[5] D. A. Norman, S. W. Draper, User Centered System Design : new prospect on Human Computer Interaction, Lawrence Erlbaum Associates Hillsdaie, NJ, 1986.

[6] R. Marom, P. Gros, "Remote teaching application : ergonomics/UI specifications", BETEL internal document, EUR-001, May 1993.

[7] XVideo Technical Overview Release 1.0, Parallax Graphics, Inc. 1991.

[8] M. Kunt, Techniques modernes de traitement numérique des signaux, Collection Électricité, Presse Polytechniques et Universitaire Romandes, 1991, pp. 175-190.

[9] SPARCstation 10 System Architecture: Technical White Paper, 1992.

[10] K. Ramkrishnan, "Performances Considerations in Designing Network Interfaces," IEEE JSAC, February 1993.

[11] Jacobson, V. Leres, C., and McCanne, S. TCPDUMP(1), 1992.

[12] JV2 Hardware Functional Specification 2.0. DEC, 1993.

[13] G. Conti, "Theoretical Performance Evaluation of Video Conferencing using the JV2 Compression Board", Technical Report, EPFL DI-LTI, 1993.

[14] D. Banks, M. Prudence, "A High-Performance Network Architecture for a PA-RISC Workstation", IEEE JSAC, Feburary 1993.

[15] P. Druschel et al., "Network Subsystem Design: A Case for an Integrated Data Path", Technical Report, University of Arizona, 1993.

[16] HP Apollo 9000 Series 700 Performance Brief, 1992.

[17] Y. Le Moan, "Effective Bandwidth at B.UNI for the LAN interconnection Service Supported by ATM Based Broadband Networks", Betel Internal document CIT-9, June 1993.