

From Remote Media Immersion to Distributed Immersive Performance

A.A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos and C. Kyriakakis
Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089
Phone: 213-740-4622

[sawchuk, echew, rzimmerm, chrisp, ckyriak]@usc.edu

ABSTRACT

We present the architecture, technology and experimental applications of a real-time, multi-site, interactive and collaborative environment called Distributed Immersive Performance (DIP). The objective of DIP is to develop the technology for live, interactive musical performances in which the participants - subsets of musicians, the conductor and the audience - are in different physical locations and are interconnected by very high fidelity multichannel audio and video links. DIP is a specific realization of broader immersive technology - the creation of the complete aural and visual ambience that places a person or a group of people in a virtual space where they can experience events occurring at a remote site or communicate naturally regardless of their location. The DIP experimental system has interaction sites and servers in different locations on the USC campus and at several partners, including the New World Symphony of Miami Beach, FL. The sites have different types of equipment to test the effects of video and audio fidelity on the ease of use and functionality for different applications. Many sites have high-definition (HD) video or digital video (DV) quality images projected onto wide screen wall displays completely integrated with an immersive audio reproduction system for a seamless, fully three-dimensional aural environment with the correct spatial sound localization for participants. The system is capable of storage and playback of the many streams of synchronized audio and video data (immersidata), and utilizes novel protocols for the low-latency, seamless, synchronized real-time delivery of immersidata over local area networks and wide-area networks such as Internet2. We discuss several recent interactive experiments using the system and many technical challenges common to the DIP scenario and a broader range of applications. These challenges include: (1). low latency continuous media (CM) stream transmission, synchronization and data loss management; (2). low latency, real-time video and multichannel immersive audio acquisition and rendering; (3). real-time continuous media stream recording, storage, playback; (4). human factors studies: psychophysical, perceptual, artistic, performance evaluation; (5). robust integration of all these technical areas into a seamless presentation to the participants.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ETP '03, November 7, 2003, Berkeley, California, USA.
Copyright 2003 ACM 1-58113-775-3/03/00011...\$5.00.

Categories and Subject Descriptors

H.5.3 [Group and Organization Interfaces]: Synchronous interaction. J.5 [Arts and Humanities]: Performing arts. H.5.2 [User Interfaces]: Auditory (non-speech) Feedback and Benchmarking.

General Terms

Measurement, performance, experimentation, human factors, information interfaces and presentation.

Keywords

Remote collaboration, music performance, real-time interaction.

1. INTRODUCTION

At the USC Integrated Media Systems Center (IMSC), we pursue research, engineering, education and industrial collaborations to develop the technologies of *immersive* and other *integrated media* systems (IMS) [1],[2]. Our vision of *immersive technology* is the creation of a complete audio and visual environment that places people in a virtual space where they can communicate naturally even though they are in different physical locations.

This paper describes a real-time and multi-site distributed interactive and collaborative environment called Distributed Immersive Performance (DIP). The DIP environment, one of IMSC's primary initiatives, forms an ideal testbed for multimedia creation, archiving, representation and transmission of electronic experiences. As a system, it facilitates new forms of creativity by enabling remote musical collaborations. Musical collaboration demands a level of fidelity and immediacy of response that makes it an ideal testbed for pushing the both the limits of human perception as well as current technology. The performance of such an experiential system can be measured and quantified through the capture, replay and analysis of digital music signals.

We first describe an initial realization of immersive technology developed at IMSC, an on-demand Internet application called Remote Media Immersion (RMI). The remainder of the paper focuses on the concepts and technical challenges of creating a Distributed Immersive Performance Environment. We describe the DIP experiments conducted in the past year, including a cello master class (asynchronous playing) and two collaborative performances (duets, synchronous collaboration), documenting the experimental setup and the musicians' experiences. Table 1 shows the timeline of events. Any musical performance creates a complex web of human relationships: between person and person (musician or audience), between individual and group, and between group and group [3]. Descriptions follow for future experiments involving Distributed Immersive Performance and

proposals for modeling and archiving the users' experiences and these human relationships.

2002	
Oct 29	<p>Internet2 Meeting: RMI Demonstration</p> <p>Performance by New World Symphony (immersive audio and video) streamed from server at ISI East cross-country to USC.</p>
Dec 28	<p>DIP Experiment 1: Distributed Duet</p> <p>A two-way duet (audio only) with violist Wilson Hsieh and pianist Elaine Chew at remote USC sites. Low latency transmissions and immersive audio creates virtual presence. at one site.</p>
2003	
Jan 18	<p>Recording from Stream</p> <p>Multi-channel audio is streamed from the New World Symphony in Miami and recorded at USC in real time.</p>
Jan 19	<p>DIP Experiment 2: Remote Master Class</p> <p>A two-way remote master class with cellist Ron Leonard at USC and students at the New World Symphony in Miami. Large screen video and immersive audio.</p>
Jun 2-3	<p>DIP Experiment 3: Duet with Audience</p> <p>A two-way performance with Elaine Chew at a piano in USC's Thornton School of Music and Dennis Thurmond with the accordion and audience at the School of Engineering. Large screen projection and PC monitor, earpiece audio and immersive audio.</p>

Table 1: Timeline of Remote Media Immersion and Distributed Immersive Performance experiments and demonstrations described in this paper.

2. RELATED WORK

There are a number of projects related to DIP being pursued by academia and industry. In a one-way non-interactive experiment in April 2000, the first live high definition television (HDTV) newscast produced over the Internet successfully reached television viewers in Seattle [4]. The demonstration was a joint effort by the University of Washington with support from several industrial partners. Other research is pursued at ATR in Japan, and several universities in the US and Europe. The European Union's Information Technologies Program [5] currently funds an ongoing multi-year project spanning several European nations on Multisensory Expressive Gesture Applications (MEGA) [6]. The research centers on the modeling, real-time analysis and synthesis of emotional content in music performance. While one of the proposal's stated goals is to create performance projects utilizing networked musical communications, at present the event descriptions do not contain evidence of networked media.

Artistic collaboration over long distance, also known as "long-distance art," is not a new concept. In 1977, video performance art for Document 6 (see [7].) was broadcast live by global satellite. In 1993, experiments with networked music performance over the Internet were done at USC's Information Sciences Institute. The network synchronization protocol [8] and examples of music that can tolerate delays can be found at [9]. The number and complexity of distributed music performances have increased through the years and include the 1998 "*Mélange à trois*", an audio-only three-way collaborative performance between Warsaw, Helsinki and Oslo, [10] and the 2002 Jam Session between Stanford's Soundwire group and McGill [11]. A list of milestones in real time networked media can be found at [12].

The problems encountered in these prior distributed performance experiments relate to the realism and immediacy of the inter-musician feedback, a requirement for effective communication in a demanding performance situation. The problems can be categorized as: (1) network delays, (2) signal synchronization, (3) echo cancellation, and (4) non-immersive audio and video acquisition and rendering. All the performances suffered some combination of the three, and the compensatory measures (if employed) include: extensive practice to modify musicians' responses; providing only local feedback and using a conductor or timing tracks to synchronize the music streams; and, performing music not based on pulse, music that can withstand random delays or music with built-in buffer zones that allow musicians to "reset" the clock. For example, in the Stanford-McGill cross-continental jazz jam that took place in June of 2002, the musicians performed with audio latencies around 90 ms (when video was eliminated), even greater video latency and no attempt was made to synchronize the two.

All these efforts differ from our DIP approach in several ways. The synchronization between multiple channels of video and audio streams (when it occurs) was either manually achieved or manually adjusted in the above mentioned experiments. However, synchronization is essential, particularly in applications that involve human observers. Immersive distributed applications are the next frontier in networked communication. To preserve the effect of immersion, participants must be provided with a sense of "immediacy", which requires that latency between participants is minimized. To achieve this level of illusion, tight synchronization between participants is required. The transmission, storage and retrieval of multi-channel audio and video while preserving both *inter-* and *intra-channel* time dependencies is an integral aspect of the DIP architecture. Furthermore, most related projects either have no provisions for a storage repository or include a solution that is special-purpose.

3. PRIOR WORK: THE REMOTE MEDIA IMMERSION PROJECT

An initial realization of the immersive technology called Remote Media Immersion (RMI) was created at IMSC for capture, storage, transmission and reproduction of audio and video presence (<http://imsc.usc.edu/rmi>). The RMI system is the result of integrated efforts across IMSC among researchers in immersive audio (Kyriakakis and Holman), data storage and streaming (Zimmermann and Shahabi) and error correction (Papadopoulos). The Corporation for Education and Network Initiatives in California (CENIC) awarded IMSC's RMI project an Honorable Mention in the "Gigabit or Bust" category in May, 2003.

To create a convincing sense of immersive presence requires the delivery of extremely high fidelity picture and sound that approaches the limits of human perception. RMI incorporates several technical innovations to achieve this:

1. acquiring video at data rates of 100 Mbps, transmitting and displaying it at data rates of 45 Mbps or more to provide high-definition (HD) or better quality.
2. acquiring and transmitting more than 16 channels of immersive audio sampled at 48 kHz and 24 bits encapsulated in a 32 bit carrier, resulting in a bit rate of 1.5 Mbps per channel. The received audio channels are processed at the client end and rendered over the 10.2 channel system developed at IMSC.
3. capturing the acoustical signature of the remote environment in advance, and then synthesizing the signal in multiple virtual microphone locations to generate a more immersive sound field.
4. devising architectures for real-time storage and playback of these multiple independent streams of video and audio data from heterogeneous, scalable, distributed servers.
5. development of protocols for synchronized, efficient real-time transmission of multiple video and audio streams from multiple distributed servers over local area and wide-area shared networks. We impose rigid bounds on time delays among the streams, latency and quality of service (QOS). The strict quality requirements are necessary to avoid glitches, hiccups, artifacts and loss of immersive realism.
6. the robust integration of all these technical areas. In addition to protocols to overcome network losses, RMI includes retransmission and other protocols at the application and perceptual layers to overcome losses in the transmission process and provide a seamless experience for the users.

In October 2002, RMI was featured at the Fall Internet2 Consortium Member Meeting hosted by USC [13]. IMSC demonstrated RMI in presenting a recorded concert performance by the New World Symphony (NWS) of Miami Beach, FL (shown in Figs. 1 and 2). The New World Symphony [14] is a post-graduate institution that educates young musicians for leadership and performance careers in orchestras and ensembles around the world, and is one of our partners in this work. At this event, the content was streamed cross-country over a shared high-speed network from a server in Arlington, VA. An audience in the 550 seat Bing Theater on the USC campus experienced the event on a large screen. An immersive sound system provided more than twelve audio channels while the video was projected in high-definition 720p format. Error correction protocols developed at IMSC were utilized to ensure seamless transmission without glitches.



Figure 1: Synchronized immersive audio and video from a concert is acquired and transmitted by cameras and an array of microphones.

Figure 1 shows RMI used in the context of capturing, transmitting and reproducing the concert. An array of microphones is placed in the concert hall (Figure 1) and the audio signals they measure are transmitted using high-speed wide-area networks (WAN) such as Internet2 or a local-area network (LAN). Simultaneously, very high resolution video (high-definition (HD) video or better) is captured by one or more cameras and also transmitted through the high-speed network. Both audio and video may also be recorded locally or at sites located anywhere on the network for off-line playback. Following this acquisition and data storage, the processed audio is reproduced (rendered) in an immersive environment that accurately preserves the audio frequencies and their correct spatial relationships to the listener, thus completely reproducing the ambience of the original concert venue. Combining synchronized high-definition video with the immersive sound completes the immersive experience (Figure 2). The reproduction site can be as small as a living room or as large as an auditorium.



Figure 2: The immersive audio and video is transmitted from a server through a shared high-speed network to a client location. There the audience is immersed in an audio environment that accurately reproduces the audio frequency content and spatial relationships of the concert. Combining the immersive audio with high-definition (HD) video completes the system.

RMI is unique and successful because we completely control the end-to-end process, from capturing the content, interfacing to the network, transmitting it without perceptual loss of information or quality, and rendering it at multiple geographically distant locations. It transforms the Internet from a low-fidelity medium for browsing information to a high-fidelity medium delivering a rich experience beyond any home medium in existence today. It is one of the few Internet applications (of which we are aware) that goes beyond the network connections and raw bandwidth requirements of an application.

On January 18, 2003, the uni-directional streaming of pre-recorded multichannel audio was extended to multichannel audio recorded over the internet. Using Papadopoulos' streaming software combined with selective retransmission we recorded an entire performance of the New World Symphony live from Miami's Lincoln Theater. The recording was done with eight channels, the number of microphones available in Miami. About two hours of material was captured with only a very small number of clicks that are due to the limitations of the network equipment in Miami.

To our knowledge this was the lowest latency remote performance ever with a latency of about 150 ms (35 ms due to the distance and about 120 ms buffers to ensure smooth playback).

4. DISTRIBUTED IMMERSIVE PERFORMANCE (DIP): THE VISION AND THREE EXPERIMENTS

Our current work is a real-time, multi-site, interactive specific realization of the immersive environment called *Distributed Immersive Performance (DIP)*. The Remote Media Immersion experiments and demonstrations primarily utilized uni-directional transmissions and off-line audio and video processing. The DIP project leverages the RMI technology and extends its capabilities to multi-directional, multi-participant and real-time interaction in synchronous and collaborative music performance. The goal of the DIP project is to develop the technology for performances in which the participants - subsets of musicians, the conductor and the audience - are in different physical locations and are interconnected by high fidelity multichannel audio and video links as shown in Figure 3.

There are generally two classes of participants with different sets of objectives and requirements. We label the participants *active* or *passive* depending on their level of interaction and the latency they can tolerate. For example, in a tele-conference application, all participants are generally active. For a performance event, there is a mix of active participants (musicians in a concert, players in a sports event, panelists at a meeting whose primary actions are those of doing, of physically engaging in the activity) and passive participants (the audience, whose primary actions are seeing and listening). Figure 3 depicts one instance of a distributed immersive performance involving a total of five sites, three player sites (each with one or more musicians), a conductor site and an audience site, all in different locations. The players respond to the hand gestures of the conductor (video signals depicted by the large arrows (large bandwidth "fat pipes")), and the conductor hears the musicians via lower bandwidth audio connections (the narrow yellow lines). In order to synchronize their playing, the musicians must be able to at least hear (and perhaps see) each other, and these lower bandwidth audio connections are also shown. The resulting concert must be

delivered in synchronism to a passive audience in another location as depicted.

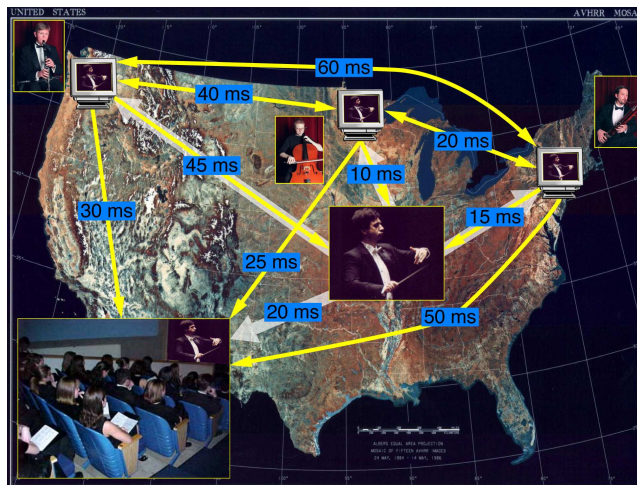


Figure 3: The Distributed Immersive Performance (DIP) concept and typical transmission latencies.

Note that musicians at live performances are able to overcome inherent audio transmission latencies among themselves, at least for distances within a concert hall. Figure 4 shows the typical audio latencies for musicians in a symphony orchestra. The speed of sound is approximately 3×10^2 meters per second, as compared to the speed of light at 3×10^8 meters per second. In this scenario with a live conductor, the visual delay of the conductor to each musician is zero. It is notable that the network latency shown in Figure 3 maps to the audio delay on stage as shown in Figure 4 and the fact that musicians can adapt to the inherent latency to play in synchronism. Thus, a major challenge in this project is to quantify and measure the psychophysical, perceptual and artistic effects of audio and video latency, synchronization, compression, bandwidth, noise, packet loss and other physical parameters on two-way and multi-way musical and other forms of human communication.

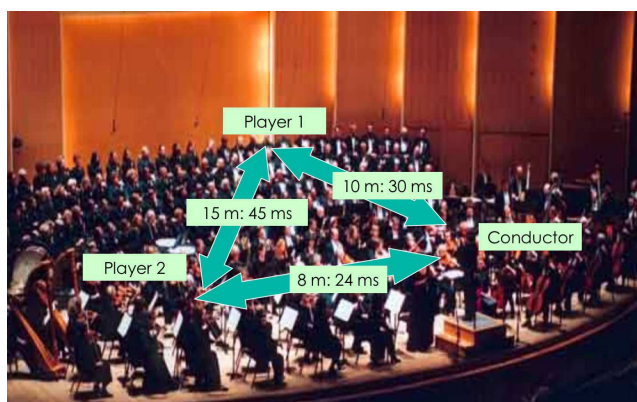


Figure 4: Audio latencies in a live orchestral performance.

4.1 DIP Experiment 1: A Two-way Interactive Duet (Immersive Audio only)

Our very first DIP experiment was a distributed piano-violoncello duet between two remote sites that took place on December 27, 2002. Pianist Elaine Chew was in the Powell Hall of Engineering (PHE)

with a keyboard and violist Wilson Hsieh was in the Hughes Electrical Engineering Center (EEB) with his viola. The players had performed together in various chamber music groups since 1992. They played selections from Hindemith's *Viola and Piano Sonata No. 4 in F major Op.11* and Piazzolla's *Le Grand Tango*. In the experimental setup, there was only a one-channel playback at the piano site, but the audio presence at the violist site (in EEB) was enhanced using the 10.2-channel immersive audio developed by Kyriakakis and Holman. To minimize the audio latency, we used a low-latency software for multichannel audio streaming developed by Papadopoulos and Sinha. The software allowed us to simulate any amount of latency during the performance by simply introducing a digital delay in the audio ranging from zero to more than 300 ms. The results were extremely promising, and we report the observations from this first experiment:

1) The latency tolerance is highly dependent on the characteristics of the piece of music and the instrumentation. For example, the first movement of the Hindemith sonata was slow and romantic, allowing greater latitude in synchronized playing. The final movement, however, on many occasions require eye-contact to synchronize sudden onsets after a pause. The players made do by listening out for each other's breathing cues, but it was impossible to replicate the results for a single-site performance. For the Piazzolla piece (a fast paced tango), when Chew used the synthesized accordion sound on the keyboard, the tolerable latency was about 25 ms. When she played the same piece with a synthesized piano sound the tolerance reached 100 ms. This could be due in part to the acoustic properties of the sound as well as her pre-disposition toward the piano.

2) Spatial sound reproduction seemed to make a huge difference for Wilson who seemed a bit weary of the experiment until we added multi-channel sound to make him feel like he was on the stage of a big hall. Hsieh noted early on in the experiment that the acoustics were "unnatural". This was after only a few bars of playing together and by simply playing back one channel at each end (the piano and the viola). He said that there was no sound coming from behind him and that made it sound "like a void". This is exactly the same result that was found in the literature in the 70's in which musicians wanted to hear music reflected from behind them (as it would from the back wall of the stage). We then used Virtual Microphones to recreate the reflected signal in the surround channels and this immediately had an effect on Wilson's acceptability of the situation. After a couple of tries we balanced it just right and he seemed to really be able to play more naturally after that.

3) Hsieh also recommended that different acoustic environments be generated at different sites. According to him, "One of the things that musicians try to do is deal with the differences in acoustics from on stage and from the audience." For some pieces, it would be good for the performers to have a very intimate sound, while for the audience it would be desirable to get the acoustics of a big hall.

4) The perception and effect of latency and hence the experience is quite different at the two sites. To properly archive a DIP experiment, the experience needs to be documented at each of the different sites. Due to the sequential nature of time, what arrives late at site A did not sound late at site B. So, for example, when the pianist anticipates and is in sync with the violist's playing, she is heard as being late to the violist. In order for the playing to be synchronized at the violist's site, the pianist actually has to play before the beat, which is precisely what a timpanist has to do in an

orchestra, guided by the conductor. In the more intimate setting of a small ensemble, there is greater autonomy and leeway for spontaneous interaction. It still remains to be seen the degree to which a distributed immersive environment can be employed for a small ensemble such as a piano trio or jazz quartet.

We collected and recorded performance sections for various delays and different pieces and will be analyzing those for some further observations. Musicians can adapt to varying delays depending on the piece and it will take a fair bit of experimentation to determine a generalizable range of acceptable delays.

4.2 DIP Experiment 2: Remote Master Class (Audio and Video)

Another DIP experiment was a remote master class experiment shown in Figure 5. On January 19, 2003, Ron Leonard, a renowned cellist from the Los Angeles Philharmonic who has taught at the Eastman School of Music, came to the Powell Hall of Engineering and conducted master classes with students from the New World Symphony for three hours. In the past, he had to fly to Miami to conduct such master classes.

We used off-the-shelf video hardware/software (Star Valley MPEG2 codecs) and virtual microphone techniques to convert (spatialize) two channel audio coming from Miami to 10.2 channels. In this experiment, we used spot microphones very close to the musicians at each venue to minimize the audio acquisition latency. The virtual microphone algorithms then synthesize the signals that would have been recorded in a larger concert hall.

Combining the Kyriakakis' virtual microphone technology with the life-size picture had created a high degree of presence. In fact, when the projector was turned off for the last 45 minutes (due to overheating) and the picture was put on a small monitor Leonard immediately commented on two things: (i) he thought Kyriakakis had turned up the volume on the sound (which he did not); and (ii) he no longer felt like the student was "really there".

A common practice in musical instruction is for the teacher to play with the student to demonstrate an interpretation of the piece. The ability to do this was limited more by the video codec latency rather than the cross-country network latency.

4.3 DIP Experiment 3: Two-Way Interactive Duet (Audio and Video)

A third Distributed Immersive Performance experiment was conducted in June 2003, this time with an audience at one of



Figure 5: A remote master class experiment.



Figure 6: The two sites of a Distributed Immersive Performance on USC's campus, June 2003.

the musician sites. The occasion was a demonstration for a National Science Foundation site visit as part of IMSC's annual evaluations by NSF. The DIP concept was demonstrated with two musicians playing a tango together while physically being present

in two different locations. The photo in Figure 6 depicts Dennis Thurmond (accordion) from USC's Thornton School of Music and Elaine Chew (piano, remotely on screen) performing together *Le Grand Tango* by Piazzolla.

Elaine Chew played the piano in USC's Ramo Hall 106 while Dennis Thurmond played the accordion in front of the audience at the Powell Hall of Engineering (PHE 106). Thurmond and the audience could see Chew on a large screen while Chew viewed Thurmond on a 20" PC monitor. At the audience site, Chew's playing was rendered using 10.2 channel immersive audio and Thurmond's accordion sounds were captured using a microphone placed close to the accordion and behind the speakers. To bypass the need for echo cancellation at the piano site, Chew wore a headpiece in one ear to hear Thurmond's playing. The two locations are separated by approximately one quarter mile on the USC campus, and were connected by the shared 100 Mbps commodity campus Ethernet. Other than the special protocols and error correction techniques that we describe in the latter part of this section, no special network arrangements were used.

The duo performed an eight minute adaptation of Piazzolla's *Le Grand Tango*. The piece was selected again because it was particularly challenging. The piece was fast-paced and filled with syncopations (off-beats). It required many expressive nuances and tradeoffs in leading the beat. Hence, it provided a good test case to explore the feasibility of distributed performances.

At 120 beats per minute (quarter note = 120 bpm) and progressing at a steady sixteenth note clip, events would happen at every 125 ms. At this pace, even a roundtrip latency of 60 ms could be debilitating. The musicians compensated by anticipating each other's actions and scaling back on the degree of spontaneity. Various checkpoints were agreed upon prior to the performance in order to anchor the synchrony. According to Chew, "By the time I heard the reaction to an action I initiated, many more notes would already have had to been played. I had to craft a musical interpretation and hold a steady pulse while keeping an ear out for delayed cues of possible ensemble issues at Powell. Dennis, on the other hand, had to anticipate my every move to make sure that the piece would be synchronized at the audience site to create a coherent performance. Once or twice during the eight minute piece, a little creative license was taken to ensure that all ends were met."

The technical details of this setup were as follows. Two channels of uncompressed, low latency audio were streamed between the two locations (16-bit PCM, 48,000 samples per second). The buffers were configured with approximately 18 milliseconds of latency (Soundblaster Live sound cards with ALSA drivers under Linux). Feedback and echo were controlled with judicious microphone and speaker placement in PHE 106 and earphones for Elaine in RHM 106. Video streaming was performed with Zimmermann and Desai's live streaming software and achieved NTSC quality (853x480 pixels at 29.97 frames per second) by capturing DV compressed video from two Canon MiniDV camcorders. The camcorders were connected to dual processor (Xeon 2.6 GHz) Linux computers via Firewire (IEEE 1394). The data was then packetized and streamed over the campus IP network to the second location of the performance. Each video stream required 31 Mb/s of bandwidth. At the receiving end, the compressed DV video was re-sequenced, decoded and displayed with a software DV decoder on the local screen. We used the 16x9 and progressive modes of the Canon XL1S camera in RHM

106 to achieve a pleasant viewing experience on the 16x9 projector screen in PHE 106, where the audience was located as well. To reduce video artifacts produced by network packet loss, missing packets were replaced rather than dropped and each video frame was kept equi-sized with 250 packets each. Therefore, missing packets affected only a single frame (approximately 33 ms) and do not cause any synchronization problems through error propagation. The end-to-end delay measured with this setup (camera to display) was between 120 to 130 milliseconds one way. Interestingly, it proved to be distracting for the performers when the audio latency was increased to match the video latency. They preferred a short audio latency and therefore the video being out-of-sync with the sound. The short audio latency was considered more critical for picking up aural cues between the performers than the video-sound synchronization

5. TECHNICAL CHALLENGES AND APPROACH

We describe here details of the technical challenges of Distributed Immersive Performance (DIP).

5.1 DIP Server, Network and Client Components

Figure 7 shows a block diagram of *one unidirectional connection* in both the RMI and DIP systems. Shown at the left of Figure 7 are server (acquisition) hardware, including an HD or FireWire DV camera and recorder. The uncompressed digital video output rates of an HD and FireWire camera are 1.5 Gbps and up to 400 Mbps respectively. This data is compressed down to 100 Mbps or less for most applications. (For comparison, note that the data rate of broadcast HDTV is 20 Mbps, and DVD video is approximately 4 Mbps). The problem with compression is greatly increased latency due to buffers that hold several frames of video (the delay between each frame is approximately 33 ms). In unidirectional applications, this processing latency is irrelevant, but for DIP interactive applications, compression with reduced latency or minimal compression may be required. Sixteen or more channels of audio are acquired simultaneously with the video and transmitted over a high-speed network.

A general network with 100 Mbps and 1 Gbps bandwidth is shown at the bottom center. Items shaded at the top center are recording, storage and retrieval (server) hardware that may be part of the local data sources or placed elsewhere on a network. A novel continuous media server architecture called *Yima* (described later) capable of scalable storage and retrieval of many synchronized streams of data in any format. For unidirectional experiments such as RMI, we first record the program content offline and stream it later from the Yima server.

The client (rendering) site shown at the right may be a long distance away from the other sites. Video from the network drives HD projectors whose output is displayed on a large perforated screen located in an acoustically isolated room. Audio from the network is mixed down to 10.2 channels and played through a speaker system surrounding the participants [3]. The speakers in the room are located to the left, center and right of the screen, and at the sides and rear of the room. The 360° immersive audio field must be precisely synchronized with the video, and the system architecture incorporates these capabilities. Dotted lines connect additional equipment (another camera and an

autostereoscopic display) needed for future stereo video. To completely interconnect N DIP sites, $2N$ sets of similar hardware are needed.

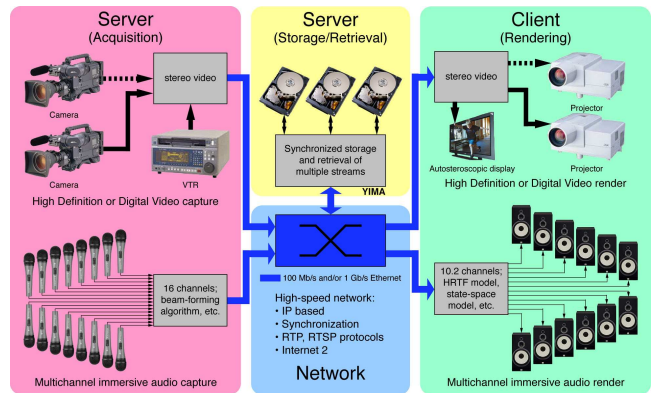


Figure 7: Fully-populated Distributed Immersive Performance (DIP) server, network and client components for a single unidirectional connection. The shaded items at the left are audio-video acquisition hardware for real-time interface to a general network, shown at the bottom center. Items shaded at the top center are recording, storage and retrieval hardware that may be part of the local data source or placed elsewhere on a network. The client (rendering) site shown at the right may be a long distance away from the other sites. Dotted lines connect optional equipment needed for stereo video. Two sets of similar hardware are needed for a full bidirectional connection between two sites.

5.2 Low Latency Real-Time Continuous Media (CM) Stream Transmission and Network Protocols

DIP poses many unique networking challenges. In addition to the high-bandwidth consumed by HD video and multiple channels of audio, DIP requires very low latency, precise synchronization and smooth, uninterrupted data flow and among many media streams in order to achieve a realistic immersive experience [15]-[18]. The single greatest limiting factor for human interaction in this immersive environment is the *effective transmission latency (delay)*. Typical latencies for the participants are shown in Figure 3. Traditional video and audio compression has been used to overcome bandwidth limitations of network transmission, at the expense of greatly increased delay. In DIP and other interactive applications, the delay due to compression may be intolerable, requiring the use of high bandwidth networks to transmit uncompressed (or minimally compressed) audio and video data [15]-[18]. Initial experiments have shown that maximum allowable latencies range from tens of to one hundred milliseconds at most depending on the experimental conditions and content.

The latency in sending acquired data through a network involves packetization, unavoidable propagation delays due to the physical speed of data transmission) queuing at each hop and processing at each hop. Our focus is on approaches to reducing queuing delays as well as on reducing the number of hops taken; this is a difficult problem as we work with best-effort networks having highly varying and unpredictable traffic patterns. The protocol stack has a relatively low delay already, and processing delays are difficult

to control, as we do not own the routers in the shared Internet2 environment.

We have assembled a test system with Linux PCs running UDP/IP and over a commodity 100BaseT network, sound cards, amplifiers, microphones and speakers. We performed several very recent experiments to determine latency effects on performances with two distributed musicians. The latency in the channel is precisely controlled from 3 ms to 500 ms, corresponding to the maximum round-trip-time (RTT) in the Internet. The communication was full duplex and the streaming software was custom-written. We found that the latency tolerance is highly dependent on the piece of music and the instruments involved. For the test piece by Piazzolla, the tolerable latency using a synthesized accordion was about 25 ms, which increased to 100 ms with a synthesized piano. We observed that musicians can adapt to some level of varying delays depending on the piece, and that spatial immersive audio reproduction made a huge difference in musician comfort by re-creating the acoustics of a large performance hall.

5.3 Precise Timing: Synchronization Using GPS or CDMA Clocks

Precise timing and synchronization of the many heterogeneous interactive streams of audio and video as it is acquired, processed and sent through a shared network to its destination is required. This implies that the latency between players must be maintained within some bounds despite the variability present in the network. In addition, musicians cannot rely on their local clocks to maintain synchronization over the entire performance, which may last hours, due to clock drift. We are developing several new transmission protocols and low-latency buffering schemes to meet these challenges. Our solution is the use of timing signals from the Global Positioning Satellite (GPS) system, which is capable of maintaining synchronization among distributed clocks with an accuracy of 10 microseconds or better. CDMA cell phone transmitter sites broadcast time signals with similar accuracy that are synchronized to GPS. Either GPS or CDMA signals may be used depending the available signal level at an acquisition (server) or rendering (client) site. In general purpose PCs, operating system delays can reach up to tens of milliseconds, which has a strong impact on the capture side, where data acquisition and timestamping occurs, and on the playback side. Our approach is to use real-time extensions to the operating system or on dedicated real-time operating systems. We are developing techniques to realign streams accurately to maintain synchronization with clock drift. During realignment, streams may be truncated or padded, and this must be done using interpolation, or repeating previous frames without introducing artifacts in playback.

5.4 Data Loss Management: Error Concealment, Forward Error Correction (FEC), Retransmission

High fidelity reproduction and accurate synchronization require a high fidelity signal, free of loss and jitter across all participants, regardless of network conditions. Packet loss is inevitable in the Internet, and strict latency requirements severely limit flexibility in error recovery. We are investigating several alternatives to this problem. Error concealment may be used in cases where network loss is sporadic without additional network overhead. We have developed multi-channel error concealment techniques using substitution and stitching that run in real-time and result in nearly imperceptible error concealment [19]. Forward error correction

offers low latency, but is susceptible to burst loss. To mitigate this problem, we pursue a multi-path streaming technique, which is a promising approach to reducing the length of a burst loss [15]. This also requires investigation of the effects of multi-path streaming on latency characteristics. Retransmission may incur unacceptable latency, especially where large distances are involved. Hybrid approaches and various blends may be possible, but add complexity. Another barrier is packet loss, which can contribute to delays (in addition to reducing fidelity), as for example during the process of reconstructing lost information. Hence, characterization of loss characteristics and methodology for dealing with such losses is an important aspect of this work.

5.5 Low Latency, Real-Time Video Acquisition and Rendering

Standard compression techniques such as MPEG are designed to transmit video at reasonable resolution over limited bandwidth networks. This is done at the expense of long latencies needed to fill a buffer with many frames for intra-frame processing. This delay may be intolerable for real-time interaction as in the DIP scenario. With the increased bandwidth available in shared local-area and wide-area networks (such as Internet2), we are exploring different parts of the compression, quality and bandwidth space to find effective techniques for DIP video transmission. For real-time, interactive applications we are investigating new low-latency compression algorithms and new types of video cameras with high speed analog-to-digital (A/D) conversion and a network, SDI (serial digital interface) or FireWire interface that outputs video compressed within a single frame (JPEG) at real-time rates (33 ms per frame). These cameras produce output at greater than 30 frames per second at QSIF (320x240) resolution and NTSC (720x480) resolution. They are close to achieving 30 frames per second at high-definition resolution (1920x1080). Current standard desktop operating systems (e.g. Windows) are not designed for such short acquisition and rendering delays, and we are developing new Linux operating system modifications with reduced latency.

5.6 Low Latency, High-Quality, Real-Time Immersive Audio Acquisition and Rendering

For accurate reproduction of audio with full fidelity, dynamic range and directionality, immersive audio requires the transmission and recording 16 or more channels of audio information, followed by processing and rendering 12 channels of audio at the client site [20]-[23]. Accurate spatial reproduction of sound relative to visual images is essential for DIP, e.g., in rendering musical instruments being played or a singer's voice. Even a slight mismatch between the aurally-perceived and visually-observed positions of a sound causes a cognitive dissonance that can destroy the carefully-planned suspension of disbelief [20],[22]. To minimize latency, we are developing new audio acquisition methods that: place microphones very close to the participants (a few meters) and reduce the analog-to-digital conversion, packetizing and transmission time to less than 10 ms. Current standard recording techniques place microphones at a longer distance to the performers such that the ambient acoustics of, for example, a concert hall are captured in addition to the direct sound from the instruments. We are working on new virtual microphone techniques to recreate the acoustics at the client with low latency [24].

Even with minimum delays in video and audio data acquisition, existing standard desktop operating systems (e.g. Windows) are not designed for short acquisition and rendering delays. As described previously, we are developing new Linux operating system modifications that provide a reduction in kernel latency. One expected result of our work is to demonstrate through subjective evaluation that the realism of immersion increases with the video fidelity and number of audio channels.

5.7 Real-time Continuous Media (CM) Stream Recording, Storage, Playback

The recording, archiving and playback of performances is essential. This requires a multi-channel, multi-modal recording system that can store a distributed performance event in real-time, as it occurs. Ideally, such a system would allow us to playback the event with a delay that can range from very short to very long. For example, an audience member who tunes into a performance a few minutes late should be able to play it while the recording is still ongoing. Hence, the system should provide at least the following functionalities: time-shifting of an event; live viewing with flashbacks; querying of streams while recording; and skipping of breaks, etc. The challenge is to provide real-time digital storage and playback of the many synchronized streams of video and audio data from scalable, distributed servers. The servers require that resources are allocated and maintained such that: (a) other streams are not affected (recording or playback); (b) resources such as disk bandwidth and memory are used efficiently; (c) recording is seamless with no hiccups or lost data; and (d) synchronization between multiple, related streams is maintained [25]-[29].

Our earlier research in scalable real-time streaming architectures resulted in the design, implementation and evaluation of *Yima* [25]-[29]. *Yima* is a second generation continuous media server that incorporates lessons learned from first generation research prototypes and supports industry standards in content format (e.g., MPEG-2, MPEG-4) and communication protocols (RTP/RTSP) as well as experimental media types.

The *Yima* server is based on a scalable cluster design. Each cluster node is a off-the-shelf personal computer with attached storage devices and, for example, a Fast Ethernet or Gigabit Ethernet connection. The server software manages the storage and network resources for DIP to provide real-time service to the various clients that are requesting media streams [25]. It provides storage and retrieval services for both HD video (MPEG-2 at 19.4 Mbps) and multi-channel immersive audio of up to 16 channels of uncompressed PCM audio samples with accurately synchronization (a total of 10.8 Mbps).

The DIP media server is based on our new High-performance Data Recording Architecture (HYDRA [27]) that extends the capabilities of the *Yima* system to include real-time stream recording (see Figure 8). For applications that require live streaming (i.e., the latency between the acquisition of the data streams and their rendering in a remote location is below a fraction of a second) the data needs to be stored on the server in real-time. Such a capability would enable digital recording, time-shifted playback, editing, pause-and-resume, advertisement insertions and more. These functionalities currently exist to some degree in single-user, consumer personal video recorder (PVR) systems such as TiVo, ReplayTV, and UltimateTV. However, these systems support only a single stream, a single media type and a single user at a time. We plan to generalize this

functionality in HYDRA as illustrated in Figure 8. Specifically we will provide support for many users and many streams concurrently [27]. Assuming a HYDRA system supports a total number of $N=n+m$ streams, any combination of n concurrent retrievals and m writes should be possible. The challenges include the design of a fine-grained locking mechanism such that the same stream can be read back after writing with minimal delay. In addition, the scheduling module of HYDRA server should be modified to support both retrieval and write threads. Write threads may be assigned different priorities than read threads. Similar to *Yima*, HYDRA supports the retrieval of many different media

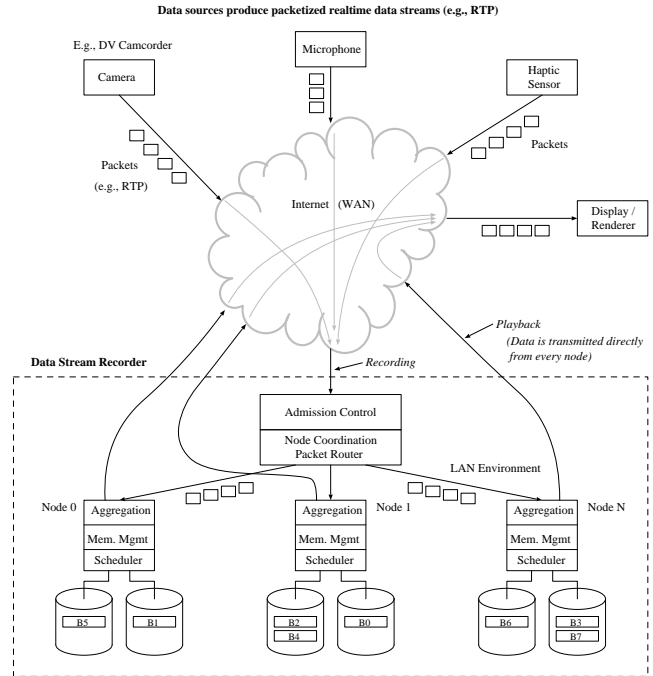


Figure 8: HYDRA recording and storage architecture.

types (e.g., MPEG-1, MPEG-2, MPEG-4, multi-channel audio, panoramic video) with various bandwidth requirements. The writing capabilities should be just as flexible and support a variety of different streams. For stored streams, flow control mechanisms can be used to regulate the optimal data flow between the source and the destination. With live stream acquisition the source must be in complete control and the destination (recorder) must absorb any variation that is provided in the data rate [29],[30]. With these extensions in place, HYDRA has the functionality necessary to support the proposed DIP experiments.

5.8 Human Factors Studies: Psychophysical, Perceptual, Artistic, Performance Evaluation

We are using the two-way interactive audio and video functions of our experimental system described in Figures 4 and 5 as a test platform to measure the effects of variable latency, feedback and echo cancellation algorithms, and various compression and error control techniques on the quantitative and perceived audio and video quality to human participants. We are making psychophysical measurements of these factors as a function of network bandwidth, compression level, noise, packet loss, latency,

etc. Algorithms for recognizing and tracking music structures will also be used to quantify synchronization among players. These methods will be used to define a metric and measure the effects of latency and presence on the music performance. The measurements are being done for various active and passive interaction scenarios, including distributed interactive musical performances with two, three or more participants, and for other types of personal interaction (meetings, lectures, etc.). The goal of these measurements is to supplement existing knowledge about these fields and measure the comfort levels of participants and musicians in two-way interaction. We also explore the minimum level of interaction needed between several musicians for effective distributed collaboration, and develop engineering metrics and parameters applicable to two-way interaction scenarios in general. We expect that some types of musical interaction among the participants may have different minimum fidelity and latency needs. For example, the interaction between the musicians at the top of Figure 2 may have lower fidelity requirements but more stringent latency requirements than the interaction between the conductor and the musicians.

5.9 Robust Integration Into a Seamless Presentation to the Participants

This challenge is the integration of research in the areas described into an experimental testbed and demonstration system. Our integrated approach is different from previous approaches and uniquely considers the entire end-to-end process of acquisition, transmission and rendering as an complete system to be jointly optimized, rather as a set of individual pieces of technology. We are integrating these technical developments into an experimental three-site DIP testbed and demonstration system for fully interconnected audio and video communication and interaction over networks. The system will test three-way interactive communications as a function of latency, bandwidth, compression level, noise, packet loss, etc. Not all sites may use the same quality of audio and video acquisition and rendering hardware in a given experiment. One possibility is the use of new network protocols that transmit side information among active participants with extremely low-latency but low data rates for musician synchronization, phrasing and coordination. In this way we test the usability of the system to participants who may have lower data rate connections. We will test minimum latency synchronized streaming of audio and video from two or more servers and make psychophysical measurements in mixed active and passive interaction scenarios. We will conduct a distributed musical performance experiments in two scenarios: with three active participants; and with two active participants and a passive audience. These experiments will further improve our understanding of the effects of fundamental communication limits on distributed human interaction. We are undertaking a complete set of psychophysical and user-centered science tests and measurements for a variety of entertainment, gaming, simulation, tele-conferencing, social gathering and performance scenarios.

6. CONCLUSIONS

We expect the DIP project to serve as a testbed for many engineering, psychoacoustical, neurological and artistic issues. It will create a forum in which electrical engineers, computer scientists, musicians and psychologists will work together in close collaboration.

Our long term goal is the creation of seamless immersive distributed environments for any type of multi-user human interaction, including entertainment, education, gaming, simulation, tele-conferencing, social gatherings, and performance events such as music, theater or sports. These applications require the development and extension of high-speed networks to provide the audio and video quality and realism for immersive interaction. Reducing delay is particularly important for interactive applications among many participants. The deployment of high-speed networks will help reduce the overall delay in transmission by simplifying the video compression and other signal processing required. Future technology may include even lower latency audio and video, more realistic and immersive video (larger screens, panoramic or hemispherical displays, 3D (stereo) video, and increased resolution). Future efforts will also be directed at archiving and re-playing the distributed immersive experience.

7. ACKNOWLEDGMENTS

We thank pianist/keyboardist Charles Dennis Thurmond, Lecturer at USC's Thornton School of Music, violist Wilson Hsieh, Associate Professor of Computer Science at the University of Utah, and Ron Leonard of the Los Angeles Philharmonic for participating in the DIP experiments. We also thank Tom Snook, Director of Internet2 Operations, his staff, and the students of the New World Symphony for their collaboration in the RMI project.

We also acknowledge the contributions of Cyrus Shahabi of the USC Department of Computer Science, and Tomlinson Holman of the USC School of Cinema-Television. The experiments could not have taken place without the assistance of numerous technical staff, including Allan Weber and Seth Scafani, and graduate students, including Rishi Sinha and Dwipal Desai.

This research has been funded by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, under Cooperative Agreement No. EEC-9529152 and Grant No. EIA-0116573. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation.

8. REFERENCES

- [1] D. McLeod, U. Neumann, C.L. Niekias and A.A. Sawchuk, "Integrated Media Systems," *IEEE Signal Processing Magazine*, vol. 16, no. 1, pp. 33-76, January 1999.
- [2] C.L. Niekias, A.A. Sawchuk, U. Neumann, D. McLeod, R. Zimmermann and C.C. Kuo, "Total Immersion," *OE Magazine*, vol. 1, no. 7, pp. 20-23, July 2001.
- [3] C. Small, "Why Doesn't the Whole World Love Chamber Music?," *American Music*, Fall 2001.
- [4] Press release. "First Live HDTV Over Internet Newscast Demonstrated at National Association of Broadcasters Convention. Las Vegas, NAB, April 2000," <http://www.researchchannel.com/special/NAB2000/press.html>
- [5] The EU's Information Technologies Program (Esprit): <http://www.cordis.lu/esprit>.
- [6] Multisensory Expressive Gesture Applications (MEGA): <http://www.megaproject.org>.
- [7] Douglas Davis: <http://sfd.com/dd/possess>.

- [8] Eve Schooler's Musical Distractions: <http://www.async.caltech.edu/~schooler/music.html>.
- [9] NOTAM: Shinji Kanki's *Mélange à trois*: <http://www.notam02.no/warsaw/melange.html>
- [10] Stanford University's SoundWire Group at CCRMA: <http://ccrma-www.stanford.edu/groups/soundwire>
- [11] McGill University's Advanced Learnware Network: <http://www2.mcgill.ca/icc/canarie/learnWare>
- [12] Jeremy Cooperstock's List of Milestones in Real Time Networked Media: <http://www.cim.mcgill.ca/~jer/research/rtnm/history.html>
- [13] Internet2 Consortium web site: <http://www.internet2.org/>.
- [14] New World Symphony web site: <http://www.nws.org>.
- [15] X. He, C. Papadopoulos, and P. Radoslavov, "A Framework for Incremental Deployment Strategies for Router-Assisted Services," *Proceedings of IEEE INFOCOM*, 2003.
- [16] S. Ghandeharizadeh, C. Papadopoulos, M. Cai, K.K. Chintalapudi, "Performance of Networked XML-Driven Cooperative Applications," In *Proceedings of Second International Workshop on Cooperative Internet Computing*, August 2002.
- [17] P. Radoslavov, C. Papadopoulos, R. Govindan and D. Estrin, "A Comparison of Application-Level and Router-Assisted Hierarchical Schemes for Reliable Multicast," *Proc. of IEEE INFOCOM* 2001.
- [18] C. Papadopoulos, G. Parulkar and G. Varghese, "An Error Control Scheme for Large-Scale Multicast Applications," *Proc. IEEE INFOCOM '98*, March 1998.
- [19] R. Sinha, C. Papadopoulos, C. Kyriakakis, "Loss Concealment for Multi-Channel Streaming Audio," *Proc. of ACM NOSSDAV 2003*, Monterey, CA, June 2003.
- [20] C. Kyriakakis, "Fundamental and Technological Limitations of Immersive Audio Systems," *IEEE Proceedings*, vol. 86, pp. 941-951, 1998.
- [21] W.G. Gardner, "3-D Audio Using Loudspeakers." Norwell, Massachusetts: Kluwer Academic Publishers, 1998.
- [22] G. Mouchtaris, P. Reveliotis, and C. Kyriakakis, "Inverse Filter Design for Immersive Audio Rendering over Loudspeakers," *IEEE Transactions on Multimedia*, 2(2), 77-87, (2000).
- [23] J. Bauck and D.H. Cooper, "Generalized Transaural Stereo and Applications," *Journal of the Audio Engineering Society*, vol. 44, pp. 683-705, 1996.
- [24] E. Cohen, J. Cooperstock, R. Zimmermann, and C. Kyriakakis, "The Challenges of Archiving Networked-Based Multimedia Performances (Performance Cryogenics)," *Proceedings of the 144th Meeting of the Acoustical Society of America*, Dec. 2-6, 2002.
- [25] C. Shahabi, R. Zimmermann, K. Fu, and D. Yao, "Yima: A Second Generation of Continuous Media Servers," *IEEE Computer Magazine*, June 2002, pp. 56-64.
- [26] R. Zimmermann, K. Fu, C. Shahabi, D. Yao, and H. Zhu, "Yima: Design and Evaluation of a Streaming Media System for Residential Broadband Services," In *VLDB Workshop on Databases in Telecommunications*, Rome, Italy, September 2001.
- [27] R. Zimmermann, K. Fu and W.-S. Ku, "Design of a Large Scale Data Stream Recorder," *Proceedings of the 5th International Conference on Enterprise Information Systems (ICEIS 2003)*, Angers - France, April 23-26, 2003.
- [28] R. Zimmermann, K. Fu, N. Nahata, and C. Shahabi, "Retransmission-Based Error Control in a Many-to-Many Client-Server Environment." *Proceedings of the SPIE Conference on Multimedia Computing and Networking 2003 (MMCN 2003)*, Santa Clara, California, January 29-31, 2003.
- [29] R. Zimmermann, K. Fu, and C. Shahabi, "A Multi-Threshold Online Smoothing Technique for Variable Rate Multimedia Streams," submitted for journal publication.
- [30] J. Escobar, D. Deutsch and C. Partridge, "Flow Synchronization Protocol," *IEEE/ACM Transactions on Networking*, 1994.