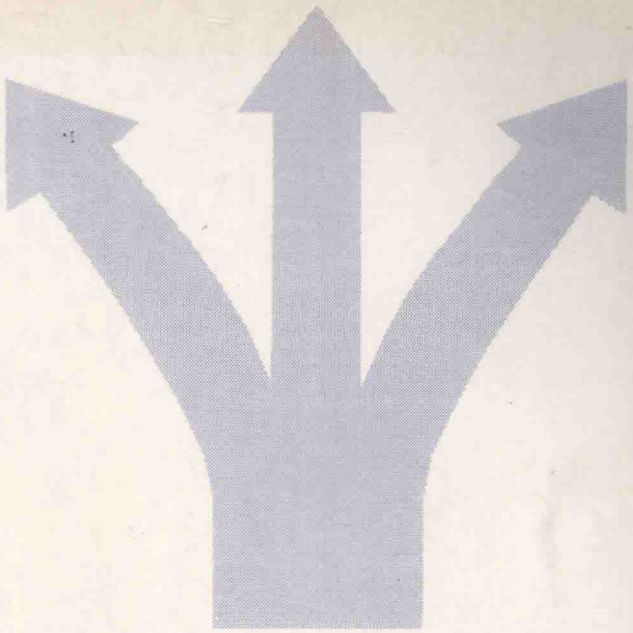
PROCEEDINGS



First International Workshop on High-Speed Network Computing

— HiNet '95 —

April 25, 1995

Santa Barbara, California

IEEE Computer Society Press

The Institute of Electrical and Electronics Engineers, Inc.

Proceedings of HiNet '95

First International Workshop on High-Speed Network Computing

April 25, 1995

Santa Barbara, California

Editors

Hussein M. Alnuweiri

Mounir Hamdi

IEEE Computer Society Press
Los Alamitos, California

Washington • Brussels • Tokyo

IEEE Computer Society Press
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1264

IEEE Computer Society Press Order Number PR07124
Library of Congress Number 95-76241
ISBN 0-8186-7124-6

Additional copies may be ordered from:

IEEE Computer Society Press
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1264
Tel: +1-714-821-8380
Fax: +1-714-821-4641
Email: cs.books@computer.org

IEEE Computer Society
13, Avenue de l'Aquilon
B-1200 Brussels
BELGIUM
Tel: +32-2-770-2198
Fax: +32-2-770-8505

IEEE Computer Society
Ooshima Building
2-19-1 Minami-Aoyama
Minato-ku, Tokyo 107
JAPAN
Tel: +81-3-3408-3118
Fax: +81-3-3408-3553

Editorial production by Robert Werner
Printed in the United States of America by KNI, Inc.

The Institute of Electrical and Electronics Engineers, Inc.

Foreword

The First International Workshop on High-Speed Network Computing (HiNet '95) took place April 25, 1995 in Santa Barbara, California. The workshop focus was on the possible impact of high-speed networks in the area of high performance parallel and distributed computing. High-speed networking is viewed by the information processing and telecommunications communities as the next major infrastructure industry. These high-speed networks, such as ATM/SONET and optical LANs, have shown great potential and benefits in various application areas including digital medical imaging, scientific visualization, video conferencing, and real-time systems.

In particular, high-speed networks have a great potential in the area of high performance parallel and distributed computing. Computers that are geographically distributed (either locally or across a wide area) can be used together to cooperatively solve problems previously requiring large and costly supercomputers. However, with this potential comes various challenges revolving around the efficient use of these high-speed networks in conjunction with the other applications areas, for example, multimedia applications.

We would like to thank all the authors and contributors as well as the members of the program committee that reviewed and commented on the submitted papers. We would also like to thank IPPS '95 chair, Viktor Prasanna, for his efforts in initiating this workshop, and Bob Werner, of the IEEE Computer Society Press, for his efforts in preparing this proceedings.

Hussein M. Alnuweiri
Mounir Hamdi
Workshop Co-chairs, HiNet '95

Workshop Chairs

H.M. Alnuweiri, Co-chair
Dept. Electrical Engineering, University of British Columbia

M. Hamdi, Co-chair
Dept. Computer Science, Hong Kong University of Science and Technology

Program Committee

S. Chanson
Univ. Science and Technology

D.H.C. Du
University of Minnesota

S. Hariri
Syracuse University

R. Jain
Ohio State University

A. Jajszczyk
Franco-Polish School

T.V. Lakshman
Bell Communications Research

W.T. O'Connell
AT&T Bell Laboratories

D.K. Panda
Ohio State Univeristy

C.S. Raghavendra
Washington State University

T. Szymanski
McGill University

Table of Contents

Keynote Address

Gb/s Networks are Here—Now!

Nanette Boden
Myricom, Inc.

Network technology, particularly that derived from interconnection technology for highly concurrent computers, has reached a point at which commodity computers can be aggregated as effectively as nodes within traditional multicomputers and multiprocessors. As an existence proof, a network technology called *Myrinet* will be described. Myrinet provides Gigabit/second links, with user-user bandwidths of several hundred Mbits/s and message latencies of tens of microseconds. The talk will address the following issues

- What will be the impact on computing of affordable Gigabit/second networks?
- New opportunities for hardware and software architectures, such as Networks of Workstations (NOWs).
- Streamlined protocols for "inter-networking" of high-speed networks.

0-8186-7124-6/95 $4.00

High-Speed Networking for Distributed Multimedia Communication

(Invited Paper)

Arif Ghafoor, Zafar Ali, Miae Woo, and M. F. Khan
Distributed Multimedia Systems Laboratory,
School of Electrical Engineering,
Purdue University, West Lafayette, IN 47907

Abstract

Considerable research in networking technology has been done over the last decade. A variety of networking infrastructures have been developed and numerous high speed networks are on the verge of deployment. These include broadband ATM networks, large scale internets and mobile communication systems. However, extensive capabilities need to be implemented in these systems in order to support sophisticated multimedia services. In this paper we discuss the potential use of these systems and discuss technological challenges. Specifically, we focus on network management and resource allocation issues. We also describe how existing and emerging networking infrastructures can be effectively used to support multimedia services.

I. Introduction

Multimedia communication is expected to be a major focus of effort in scientific and technological development during the rest of this decade. It is already opening up new research frontiers in a number of areas such as computer networks, storage systems, semiconductor technology, distributed systems, parallel processing, information theory, databases, real-time operating systems, computer graphics, human-computer interaction, algorithms, hypertext and hypermedia, etc. As a result of concerted effort in these areas, many multimedia applications involving different media types, e.g., video, audio, text, images, animation and graphics, are expected to be available in the near future. Most of these applications/services will use some form of *pre-orchestrated* information stored at various sites [10], [17], [5]. For example, various government organizations in the US have recently undertaken many initiatives to build digital multimedia libraries that will promote affordable remote learning environments and allow users to author, store and share multimedia documents interactively. In another application area known as "telemedicine", the emerging *Broadband Integrated Services Digital Network* (B-ISDN) will allow the development of medical communication systems capable of delivering medical services to distant communities as well as remote access to patient data. In the entertainment industry, several joint ventures aimed at developing interactive multimedia entertainment (e.g., video-on-demand, interactive games, etc.) have been formed in recent months [14]. Plans also include to provide services such as tele-shopping, news-on-demand and other convenient and financially viable home information services.

Typically, presentation of any pre-orchestrated multimedia information requires real-time delivery of some kind of *multimedia documents* (MMDs) to the end user. An MMD is assumed to be logically subdi-

0-8186-7124-6/95 $4.00

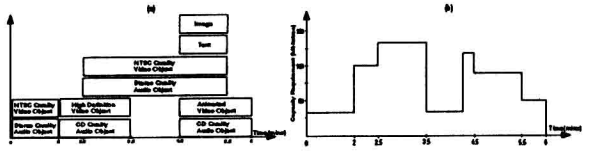


Fig. 1. (a) Time line representation of an MMD, (b) Capacity use profile of the MMD in (a).

vided into *multimedia objects*. These objects may be *continuous* media like video and audio or they may be *discrete* media objects, e.g., text and images. These objects have certain temporal relations specified at the time of the composition of the documents. Fig. 1(a) illustrates such a relation on a time line. At the time of transmission of an MMD, the delivery of its component objects must be sequenced in such a way that these objects can be played out according to the specified temporal constraints.

MMDs are assumed to be stored at some remote sites. Their retrieval requires an efficient and flexible transport mechanism. Integrated networks based on the *Asynchronous Transfer Mode* (ATM) provide a flexible transport service with high utilization of network resources [11], [13]. ATM is a packet-oriented transfer mode based on asynchronous time division multiplexing and all information is conveyed using fixed sized packets called *cells* [12]. Bandwidth efficiency in ATM networks is increased by statistical multiplexing of multimedia traffic at the expense of cell-to-cell delay variations and cell losses. However, cell losses are guaranteed to be consistent with media-specific reliability requirements [12]. Parameters such as bit error rate and cell loss rate can be used to specify these requirements [7]. Furthermore, the International Telecommunication Union (ITU, formerly CCITT) has recommended ATM as the underlying transport mechanism for the B-ISDN [2].

For delivery of MMD over networks, it is necessary to: (i) maintain continuity in delivery of individual component objects; and (ii) to preserve the predefined temporal relationships among them. These two requirements are known as *intra-object* and *inter-object synchronization* problems, respectively. Intra-object synchronization mechanisms are required to smoothen the delivery of multimedia objects and to minimize the impact of delay jitter over the presentation process. On the other hand, inter-object synchronization mechanisms are needed to ensure correct presentation of concurrent data streams as they are communicated over different channels. Maintaining *lip-sync* [17] between audio and video objects is an example of inter-stream synchronization requirement.

As mentioned earlier, a typical MMD is composed of several media objects with different throughput requirements. Furthermore, the level of concurrency can change over the period of presentation. Thus network capacity requirement may change dramatically over the lifetime of the connection established for retrieval of MMD. This is exemplified in Fig. 1(b) where the *capacity use profile* for the MMD of Fig. 1(a) is plotted. This large variation of capacity requirement poses a new set of challenges for high-speed networks, including managing virtual channels

and designing real-time multimedia synchronization protocols. The objective of this paper is to highlight these challenges and to discuss possible approaches to solve these issues. Specifically, our discussion focuses on the following issues:

- For an MMD, how can we specify the presentation quality to help characterize the presentation process?
- How can resource requirements be established for retrieval of MMDs over high speed networks? These questions center around the issue of network resources (bandwidth and destination buffering) required to guarantee synchronized presentation of MMDs. Trade-offs between the desired performance (quality of presentation) and required network resources need to be considered.
- What are the technical issues for supporting multimedia services over the current and emerging high speed networking infrastructures? We briefly elaborate three telecommunication infrastructures that are expected to play a key role in providing synchronized MMD services to distributed users over broadband networks. These include ATM networks with multi-channel configurations, high speed internets, and mobile communication environments.

II. Synchronization Requirements of Multimedia Information

Typically, a multimedia user retrieves MMDs interactively by browsing through an information space. This interaction can be represented by a bi-level hypermedia paradigm with vertices pointing towards various MMDs and edges representing the logical links among documents. These links are generally anchored at their departure point to provide the user with some explicit MMD to activate in order to follow the link. A user can explore or skip individual nodes, thus limiting the information viewed.

Once an MMD is selected by the user, its synchronous playback requires enforcing inter-object temporal relationships among component media objects within the selected MMD. Several models have been proposed for specifying temporal constraints within an MMD, e.g., Hytime [1], Petri-net based models [9], [10], MODE [4], G-Net [6]. Most of these models, however, have one major limitation: they cannot be used directly to specify network synchronization requirements. In these models temporal specification is provided at the object level where objects can be of variable duration, such as a video clip, an audio segment, or an image sequence of arbitrary length. Synchronization at the object level is rather coarse and hence is difficult to perform synchronized transmission over the networks using these specifications. In a networked environment, synchronization needs to be performed at a finer level. One

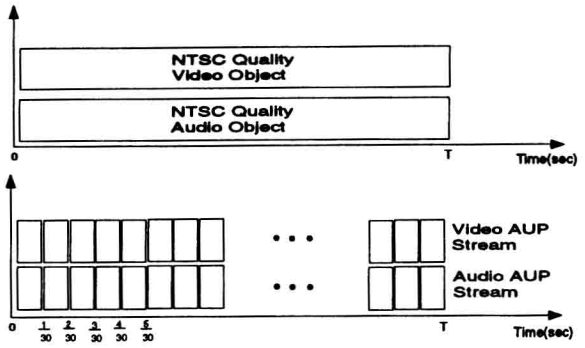


Fig. 2. Temporal segmentation of multimedia objects into AUPs.

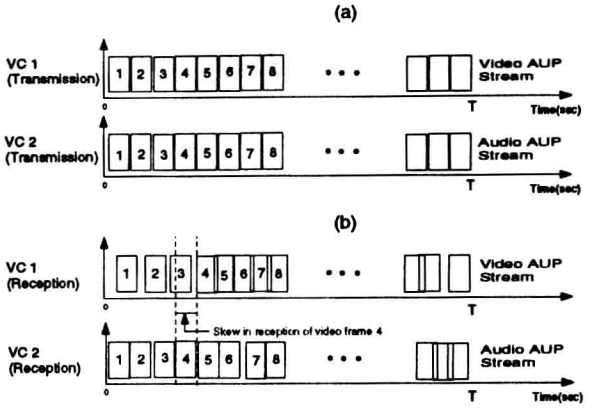


Fig. 3. Skew caused by late arrival of video frames, (a) Controlled transmission of AUPs of video and audio objects over independent channels, (b) AUPs reception at the destination.

approach is to assume that each object in an MMD can be decomposed into a sequence of *Atomic Units of Presentation* (*AUPs*) whose transmission is controlled for synchronization purposes. For example, a frame can be considered as an *AUP* for video objects, while an ATM cell containing an equal number of PCM code samples can be taken as an *AUP* for audio objects. For a discrete media object, such as an image or a text, the entire object can be viewed as an *AUP*. The transmission of an object then consists of a sequence of AUPs, with each AUP marked with an identification number. To facilitate inter-object synchronization, all concurrent objects in an MMD can be divided into AUPs of equal duration as illustrated in Fig. 2. In this figure, the duration of each AUP is chosen to be $\frac{1}{30}th$ of a second, which is the length of a video frame.

For a successful playback of an MMD, the source needs to maintain a controlled transmission of AUPs over available *Virtual Circuits* (VCs). However, the arrival time of each AUP may not exactly follow the same pattern due to the nondeterministic behavior of the network. Jitter delays may vary from AUP to AUP and may also vary among multiple VCs. This is illustrated in Fig. 3. This figure shows that due to jitter delays in the network some *AUPs* do not arrive in time. To maintain the continuity of presentation of the whole object, playout of arrived data must continue even if *AUPs* belonging to some other related objects within the MMD do not arrive. Late *AUPs* can be played out with the ongoing presentation, provided that the skew caused by the late arrivals is within some tolerable limit. For example, consider a situation shown in Fig. 3(b). AUP_3 of the video object is late in meeting its deadline by $\frac{1}{30}$ seconds. It can be played back with the current audio segment (AUP_4) that causes a skew of $\frac{1}{30}$ seconds, provided it is acceptable.

The synchronization problem becomes more acute when network resources are constrained. This is because in a resource constrained environment, main-

taining controlled transmission of AUP streams as shown in Fig. 3(a) is not possible. Determining a feasible *transmission schedule* to guarantee synchronization in spite of jitter delays and resource constraints is an important networking challenge.

III. Network Resource Management Issues

From the point of view of network resource requirements, the issues involved in retrieving MMDs over ATM networks are quite different from the ones faced in transmitting single media objects. Specifically, for retrieving a single object, the VC used for transmission can be tailored to suit its bandwidth requirements. Usually, there is a large variation (over the lifetime of the connection) in the network capacity required for retrieving an MMD. One of the possible strategies is to allocate network resources depending on peak requirements. Clearly, this can result in an unacceptably low utilization of scarce resources. The other obvious strategy is to open independent VCs for individual objects. This can result in excessive network management overheads. Furthermore, it may not be possible to open the required VCs because of unavailability of resources at a later stage. A third possibility is to use a single VC for the transmission of multiple objects.

Generally, a network connection can be a collection of VCs used in the transmission of an MMD. A typical connection of this kind is shown in Fig. 4. Note that the individual VCs in an MVC connection may follow different paths in the network.

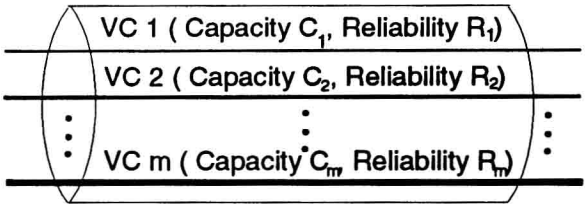


Fig. 4. A network connection consisting of m VCs.

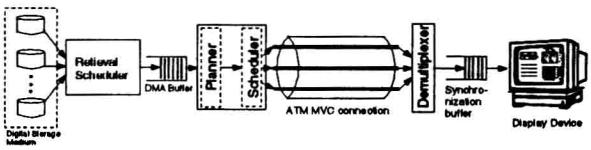


Fig. 5. MMD retrieval over a resource deficient connection.

A VC whose capacity is greater than or equal to throughput requirements of a multimedia object, and whose reliability characteristics are also better than or equivalent to the reliability requirements of the object is termed as a *resource sufficient VC* [16]. In other words, a connection that can provide an independent resource sufficient VC for all concurrent objects of an MMD is a resource sufficient connection for that MMD. That is, in a resource sufficient connection, multiplexing of concurrent object streams onto a single VC is not required.

A connection that is not resource sufficient is termed as *resource deficient*. It eliminates the above mentioned disadvantages associated with a resource sufficient connection by multiplexing several objects over the same VC. Furthermore, channel capacity utilization can be increased by pre-fetching during periods of low capacity demand.

Supporting a resource deficient connection requires efficient *scheduling mechanisms* and *resource reservation protocols* that take into account the structure of

the MMD and type of VC used for retrieval. This is illustrated in Fig. 5. As can be seen from this figure, the process of MMD retrieval consists of two steps. As part of the first step, a *resource planner* determines the number of VCs to be used, the capacity of the individual VCs, and the size of destination buffers. In practice, the planner is responsible for managing network resources in an optimal manner so that the network can support a large number of users. Based on the decision made by the planner, a *MMD scheduler* is responsible for the provision of synchronous playback of the MMD. This requires pre-scheduling the transmission of AUPs involved in the presentation, depending on the capacity of the connection and jitter characteristics of the individual VCs. The idea is to carefully orchestrate the transmission schedule at the source based on resource availability and temporal constraints specified by the playout schedule.

The network resource planning and MMD scheduling decisions are based on network states, i.e., the availability of network resources, the structure of the MMD and the desired quality of presentation. In order to make these decisions, it is important to quantify QOP of MMD's and to identify its relationship with network resource usage. Before discussing resource planning and scheduling problem, we need a set of parameters, that can be used to express preferred, acceptable and unacceptable levels of tolerance in presentation of an MMD. These parameters can also be used to determine the network resource requirements as well as to establish trade-offs between QOP and network resources.

IV. What is Quality of Multimedia Presentation?

Quality of MMD presentation that can be supported is directly dependent on the available network resources. Also, it is affected by the jitter characteristics of the network. We can define two sets of QOP parameters, one set quantifying the effect of jitter and other describing its dependence on resource availability.

Jitter related parameters deal with the deadline missing scenarios as well as buffer overflow problems at the destination. In order to reduce the chances of missing deadlines, pre-scheduling strategy [10] can be used. In this strategy, for an AUP to meet its deadline at the destination, sufficient time is allowed to overcome delay jitter. This time, called the *control time*, is indeed an artificial delay introduced in the presentation process in order to smooth out the effect of jitter on the process.

Other jitter related QOP parameters include the maximum delay that the user can tolerate between presentation of successive *AUPs* and the maximum tolerable skew among related streams. Synchronization protocols at the destination site are needed to enforce compliance of delay tolerance [15].

The second set of QOP parameters is due to the effect of network resource constraints. The parameters include latency in the start of the presentation

process and percentage of dropped AUP's due to resource limitations. Most of the multimedia applications are not too sensitive to AUP losses. Occasional loss of AUPs can be tolerated without seriously affecting the QOP. Hence, some AUPs can be dropped at the source in case resources are seriously constrained.

The QOP parameters discussed here can be used in network resource planning and for evaluating trade-offs involved in retrieving MMD over networks.

In practice, decisions regarding transmission scheduling, network management, and resource usage may not be completely independent. The goal is to maximize network resource utilization without affecting the quality of presentation. In case of severe resource constraints, the objective is to provide the best possible quality of presentation. In practice, a balance must be struck between the QOP and resource utilization. For example, high capacity connections can reduce latency while on the other hand, decreasing the allocated capacity generally increases both the delay and the size of application synchronization buffers. Fig. 6 depicts these trade-offs for the example of the MMD of Fig. 1. Although this figure is plotted explicitly for the MMD of Fig. 1, it is intended to provide a general notion of 'schedulability' for any MMD transmission. In this figure, any point on the surface $\mathcal{S}$ represents the minimum amount of capacity and the minimum size of the synchronization buffers required at the destination site that ensures transmission that maintains the desired QOP.

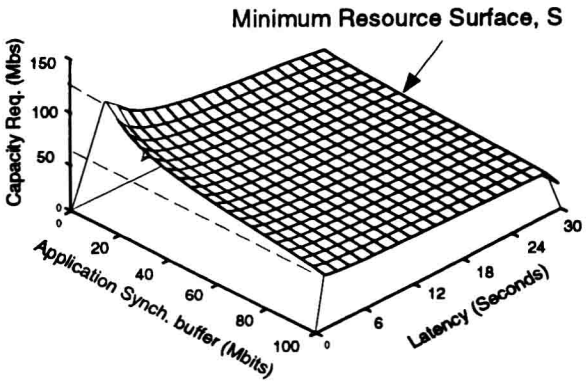


Fig. 6. Trade-offs in resource planning: Minimum resource surface for MMD of Fig. 1 (a).

In other words, this cut-off surface indicates that the desired QOP can not be met if the network resources are severely constrained. The region below the surface represents the 'unschedulable' region. The size of this region depends on the overall structure of the MMD, the number of VCs supporting the end-to-end session, and their reliability characteristics. These dependencies are discussed in detail in [3].

Such trade-off curves can be used for identifying network resources needed to support the service satisfactorily. For this purpose, a suitable characterization of the minimum resource surface and algorithms for finding optimal operating points on the surface are required. Since, resource planning needs to be done in real-time, fast algorithms would be needed. These algorithms can use *a priori* knowledge of the "capacity usage profile" of MMD and the acceptable degradation in QOP.

V. Future Networking Infrastructures

We envision that the emerging high-speed networking environments can provide tremendous communi-

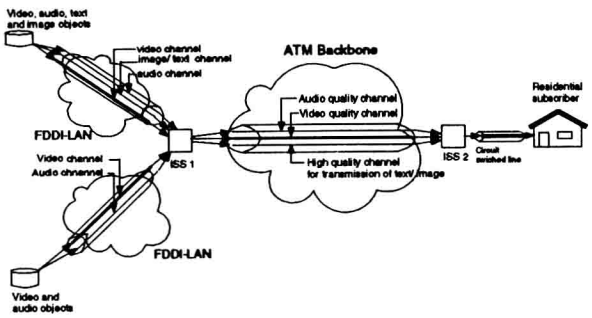


Fig. 7. Multimedia communication over a high speed internet: a distributed paradigm.

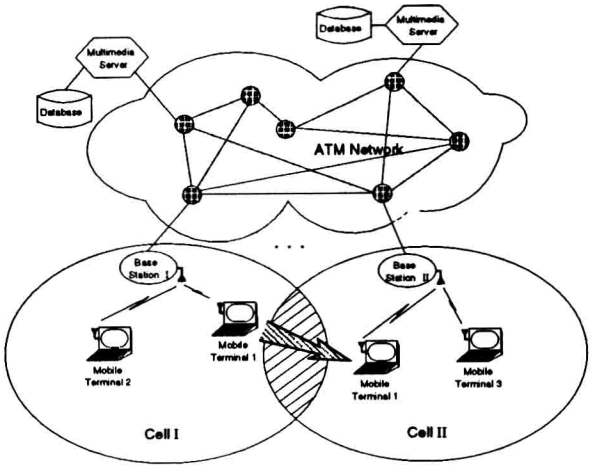


Fig. 8. A view of mobile internetworking.

cation resources. However, in order to support multimedia services over these environments, unique challenges are posed since the characteristics of the VCs used for the transmission of MMDs may be quite different from subnet to subnet. Each subnet may have its own management strategies to handle diverse media traffic, as well as its own resource constraints. Furthermore, end-to-end delay over independent virtual channels may also vary widely, requiring extensive buffering at the destination to maintain both intra-object and inter-object synchronization. Such buffering can be prohibitively large [8]. An example elaborating this scenario is shown in Fig. 7.

In this scenario, it is assumed that the MMD requested by the user is partially stored at distributed servers interconnected over an internet with an ATM backbone. An MVC connection is used for the transmission of the MMD. In order to support multimedia applications in such environments, new transport techniques will be needed.

Future networks are also expected to provide multimedia services to the mobile users. In such an environments, various multimedia servers, connected over land-based networks, will be accessible to mobile users, as shown in Fig. 8. Upon receiving a request, a multimedia server retrieves the requested MMD from its databases and communicates it over an MVC ATM connection to a base station which ultimately transmits the data to the mobile user using an RF channel as illustrated in Fig. 9.

We expect that the base station will serve as an interface between the land-based and the wireless networks, as show above. Consequently, the base station needs to perform an internetworking function for protocol conversion, to provide necessary signaling and to convert data into a suitable form. For interoperation with ATM networks, the internetworking function requires buffering capability for necessary rate adaptation, clocking adjustment and bit alignment across the boundaries between the heterogeneous networks, to name a few.

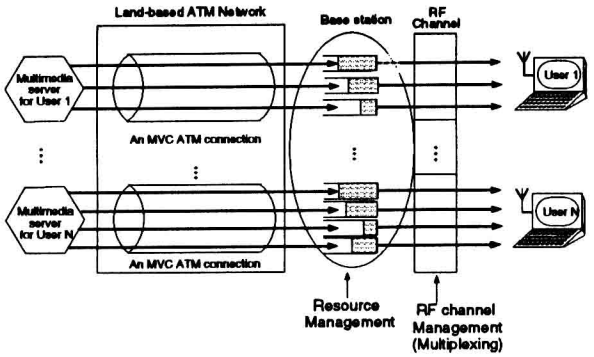


Fig. 9. An abstraction for land-based and mobile internetworking

VI. Conclusion

In this paper an attempt has been made to identify the technical issues involved in retrieving multimedia documents over networks. In particular, we discussed channel and capacity management issues. A set of quality of presentation parameters have been presented to help the study of trade-offs between the system resources and user requirements. Furthermore, Future challenges for the emerging networking infrastructures have been identified.

References

[1] Hytime: Information technology - hypermedia/time-based structuring language. International Standard ISO 10744, International Organization for Standardization (ISO), 1992.

[2] Broadband aspects of ISDN. International Standard XVIII-R 34-E, Recommendation no. I.121, CCITT Study Group COM, June 1990.

[3] Ali, Z., Woo, M., and Ghafoor, A. Delivering interactive multimedia docoments over ATM networks. Tech. rep., School of Electrical Engineering, Purdue University, 1994.

[4] Blakowski, G., Huebel, J., Langrehr, U., and Muehlhaeuser, M. Tool support for the synchronization and presentation of distributed multimedia. *Computer Communication 15*, 10 (December 1992).

[5] Campbell, A., Coulson, G., Garcia, F., and Hutchison, D. A continuous media transport and orchestration service. In *Proceedings of SIGCOMM* (Augest 1992), ACM, pp. 99–110.

[6] Deng, Y., and Chang, S. K. A framework for the modeling and prototyping of distributed information systems. *International Journal of Software Engineering and Knowledge Engineering 1*, 3 (1993), 203–226.

[7] Ferrari, D. Client requirements for real-time communication services. *IEEE Communication Magazine 28*, 11 (1990), 65–72.

[8] Ferrari, D. Distributed delay jitter control in packet-switching internetworks. *Internetworking: Research and Experience 4*, 2 (March 1993), 1–20.

[9] Little, T., and Ghafoor, A. Synchronization and storage models for multimedia objects. *IEEE Journal on Selected Areas in Communications 8*, 3 (April 90), 413–427.

[10] Little, T., and Ghafoor, A. Multimedia synchronization protocols for broadband integrated services. *IEEE Journal on Selected Areas in Communications 9*, 9 (December 1991), 1368–1382.

[11] Minzer, S. E. Broadband ISDN and Asynchronous Transfer Mode (ATM). *IEEE Communications Magazine 27*, 9 (September 1989), 17–57.

[12] Partridge, C. *Gigabit Networking.* Addison-Wesley, 1994.

[13] Prycker, M. D. ATM switching on demand. *IEEE Network 6*, 2 (March 1992), 25–28.

[14] Ramanathan, S., and Rangan, P. V. Architecture for personalized multimedia. *IEEE Multimedia 1*, 1 (Spring 1994), 37–46.

[15] Ravindran, K., and Bansal, V. Delay compensation protocols for synchronization of multimedia data streams. *IEEE Transactions on Knowledge and Data Engineering 5*, 4 (August 1993), 574–589.

[16] Woo, M., and Ghafoor, A. Multichannel scheduling for communication of pre-orchestrated multimedia information (homogeneous channels case). In *Proceedings of INFOCOM '94* (Toronto, Ontario, Canada, June 1994), IEEE Computer and Communications Society, pp. 920–927.

[17] Woo, M., Qazi, N. U., and Ghafoor, A. A synchronization framework for communication of pre-orchestrated multimedia information. *IEEE Network Magazine 8*, 1 (January/ February 1994), 52–61.

Can Shared Access Networks Adequately Support High-Performance Computing? Re-visiting MAC Protocols

William T. O'Connell
AT&T Bell Laboratories
Murray Hill, New Jersey
wto@research.att.com

Abstract

Emerging high-speed networks are broadening the arena for high performance parallel and distributed computing applications. Research is now exploring the capabilities of Sonet/ATM technology due to the in-effectiveness of traditional Medium Access Control (MAC) protocols when used with fine granularity communication. With out placing a strong level of order on medium access, chaos results with poor medium efficiency and capacity use. While under heavy loads, most medium access protocols buckle when large number of nodes try to transmit simultaneously, resulting only in thrashing on the medium. This paper will re-visit common MAC protocols comparing them against criteria required by high-performance message passing. It will then introduce concepts of a new protocol called DQRAP. This MAC protocol developed at Illinois Institute of technology has been shown not only to have a throughput of 1 under loads greater than 100%, but that there is a predictable delay from the time it takes to resolve a large number of simultaneous collisions to the time the data involved in the collision starts transmitting on the medium.

Index Terms - DQRAP, Medium Access Control, Collision resolution, high-speed networks.

1 Introduction

The desire to produce scalable parallel and distributed computing applications has focused technology towards high-speed internetworking. Over the last decade the industry has seen a growing interest in connecting two or more uni/multi-processor machines through both Local and Wide Area Networks (LANs, WANs, respectively). The existence of implementations such as PVM, MPI, Linda and D-Memo have made this possible through levels of abstractions [1]. These extended environments provide a type of processor interconnection that falls under the nomenclature "distributed-memory parallel machine", or should we say virtual parallel machine.

With this extended virtual machine, we lose certain guarantees that we would normally see in a closely coupled multistaged network, such as a Giga-switch. Referencing cable television criteria [2], I am stating that this criteria can be extended to high-speed switches. Parallel processes must have certain guarantees with respect to inter-process communication:

- predictable delay - Each transmission has a predictable delay. This does not refer to transmission latency, but packet delay before transmission starts.
- distributed control - Each node manages its own requests and transmissions independently (no master station is required for network).
- immediate access - If there is an available non-busy channel, then immediate transmission is performed.
- full channel utilization - Throughput of the system is equal to the offered load (traffic) up to a load of one, and maintains a throughput of one even when the offered load is greater than one.
- fair access - Each transmission request is met on a first-in, first-out (FIFO) basis. If priority schemes are used, protection against indefinite postponement is needed.

No matter what the available medium bandwidth is, treating nodes in a LAN/WAN interconnect as a multi-processor machine will sacrifice several to *all* of these guarantees. This is especially true with the use of traditional Medium Access Control (MAC) protocols.

Recently, the focal point has moved from traditional MAC network interfaces to Asynchronous Transfer Mode (ATM) switches. However, ATM still does not satisfy the guarantees of predictable delay and distributed control. Once loads reach certain thresholds[a] packets will be buffered and possibly lost under heavy loads. For example in a LAN scenario, when loads are sustained for a period of time at a load greater than one, flow control and buffering problems occur. This happens after a node has forwarded a packet which prevents it from *i)* buffering the packet instead of the switch doing so (which would lead to smaller buffering requirements) *ii)* calculating the effective delay (due to buffering done non-locally) and conditionally selecting not to send the packet if the delay would be greater than a threshold. This is an important consideration for time critical communication, such as continuous media feeds or real-time applications. In addition to the predictable delay problem, distributed control is partially moved to the central switch (or switches). Much buffering is now done centrally versus on

a. ATM switching is based on a statistical model of a ratio of multiplexed switched capacity to total maximum input.

0-8186-7124-6/95 $4.00 © 1995 IEEE